UNIVERSIDAD AUTÓNOMA DE MADRID

Departamento de Biología Molecular

Facultad de Ciencias

Estudio computacional de las bases moleculares de la especificidad funcional en familias de proteínas

TESIS DOCTORAL

Memoria presentada para optar al grado de Doctor en Ciencias por:

Antonio Rausell de Frías
Centro Nacional de Investigaciones Oncológicas (CNIO)
2011

Dirigida por Prof. Alfonso Valencia Herrera

A mis padres

Agradecimientos

Quiero agradecer a mi director Alfonso Valencia la confianza que ha puesto en mí a lo largo de estos años, primero dándome la oportunidad de formarme en su grupo y, de forma continuada, con las responsabilidades y retos que me ha ofrecido. Alfonso ha sido siempre receptivo a mis intereses e iniciativas y me ha motivado constantemente. Es realmente un privilegio haber podido trabajar con él de cerca y contar con su buena guía.

Además de con Alfonso, esta tesis está hecha codo con codo con David de Juan. David me ha dedicado incontables horas con una generosidad fuera de lo común, ilustrándome desde en los asuntos más elementales del día a día en el laboratorio, hasta en las ideas más complejas y potentes. Trabajar con él es "science in action" y ha logrado inculcarme la duda constante, el trabajo minucioso y la pasión por comprender.

Junto a Alfonso y David, Florencio Pazos ha participado estrechamente en todo mi trabajo, desde mi primer día en el labo cuando me hizo un hueco a su lado en aquel zulo del CNB. Sito ha estado siempre disponible y ha sido importante contar con su visión en los momentos clave.

Quiero agradecer las aportaciones directas a esta tesis de Gonzalo López, por su ayuda con el manejo de la base de datos FireDB, y a Ana Rojas y Ángel Carro por su trabajo con el servidor Treedet.

Mi formación y mi trabajo en el grupo de Alfonso durante estos años ha contado con la ayuda impagable de todos los miembros del grupo de Biología Computacional y Estructural y de la Unidad de Bioinformática (los que están y los que ya se fueron): Víctor, Txema, Jan Jaap, Ángela, Gonzalo L., Iakes, Michael, Dani, Belén, Gonzalo G., Ramón, Ana, Alonso, Pisano, Ángel Edu, José María, José Manuel, Osvaldo, Andrés, Willy, César, Paolo, Raquel, Cristina, Mark, Jorge, Gema, Leticia, Milana, Florian, Martin, Miguel, Luis S., Tamames, Gloria, Anaïs, Ashish, Guille, Adrià, Almudena, Alfredo, Allan, Christian (si se me escapa alguien que me perdone…). Muchas gracias a todos, de verdad, compartir trabajo y ocio con vosotros ha sido lo mejor de estos años y me lo llevo para siempre.

Por último quiero agradecer a mi familia y mis amigos (en especial Jaime y Víctor) su apoyo incondicional y, muy muy especialmente, a Natalia, que me ha estado animando y ayudando minuto a minuto durante la escritura de esta tesis y seguro que es quien más va a disfrutar de que la acabe: ¡gracias amor!

Summary

Throughout evolution, homologue proteins diverge in sequence as a result of the evolutionary pressure exerted on the variability arising from mutation, duplication, speciation and deletion events. At evolutionary scale, this divergence process is usually translated in an internal organization of the proteins within the family into protein subfamilies. This organization assumes that the different subfamilies represent functional features that are specific within the context of the common function of the families. Several computational studies have shown the relationship between subfamily structure, residues that are differentially conserved among the subfamilies (SDPs) and key aspects of functional specificity.

The main goal of this thesis is to deepen the current understanding of the functionally driven divergence of protein families by performing a novel study that, for the first time, analizes at large scale the relationship between subfamilies and differential interaction patterns among homologue proteins, while also taking into account the implication of SDPs in protein-protein interfaces. Consequently, this study combines the implications of functional sites with those of ligand interacting sites.

In order to perform this large-scale study, a novel sequence-based computational method to analize protein families was developed which is able to discern both the subfamilies' internal structure and their differentially conserved residues in a coherent and simultaneous manner: the S3det method. This method provides a methodology that can be applicable to a big set of proteins making it possible to obtain a representative number of subfamilies and SPDs.

The results obtained show that protein family organization into subfamilies answers in a general way to functional differential features that are both related to the specific enzymatic activity and to distinctive sets of interacting proteins. Moreover, positions that are differentially conserved in subfamilies (SDPs) appear to be structurally associated to functional regions that correspond to catalytic sites, ligand union sites and protein-protein interfaces. Indeed, such associations occur both in terms of the spatial distance distributions and in the relative enrichments. Most importantly, the implication of the SDPs in protein-protein interfaces is especially clear in the case of heterocomplexes interfaces. These observations allows to propose that binding specificity evolves by selecting key residues differentially conserved in the subfamilies as pivotal points indicative of binding with their effectors.

In a complementary manner, two other computational methods were developed (Xdet and Mcdet). The former exploits quantitative functional information while the latter makes use of supervised classifications, and both are used to predict residues determining functional specificity. Their predictive ability has been demonstrated in protein alignments where sequence similarities do not correspond to functional similarities.

Taken together, the results of this thesis provide generality and quantitative support to the hypothesis stating that sequence divergence accumulated in protein families is driven by functional divergence.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ABREVIATURAS

MCA: Análisis Múltiple de Correspondencias (del inglés *Multiple Correspondence Analysis*).

MSA: Alineamiento Múltiple de Secuencias (del inglés *Multiple Sequence Alignment*).

PCA: Análisis de Componentes Principales (del inglés *Principal Component Analysis*).

SDP: Posición Determinante de Especificidad (del inglés *Specificity Determining Position*).

I. Introducción

I.1. De la era genómica a la post-genómica

En 1995 se publicó por primera vez el genoma completo de un organismo, el de la bacteria causante de la meningitis *Haemophilus influenzae* (Fleischmann *et al*. 2005). Con una tendencia de crecimiento exponencial, quince años después el número de genomas totalmente secuenciados se acerca ya a los 900, según las estadísticas recogidas en la base de datos GOLD (*The Genomes On Line Database*, Liolios *et al.* 2008). A lo anterior se suman los proyectos metagenómicos y de secuenciación medioambiental (secuenciación de ecosistemas enteros sin consideración de organismo), que están incrementando considerablemente la cantidad de secuencias de proteínas disponibles (Venter *et al*. 2004; Yooseph *et al*. 2007). El repositorio general de secuencias de proteínas es UniProt (Apweiler *et al.* 2004). En su componente supervisada de forma manual (UniProtKB/SwissProt, versión 57.1 de 5 de abril de 2011) reporta ya 526.969 secuencias, mientras que UniProt/TrEMBL -resultante de la conversión automática de secuencias de nucleótidos procedentes del EMBL- contiene 14.555.721. Por otra parte, los proyectos de genómica estructural (Sali 1998; Brenner 2001) están resolviendo estructuras tridimensionales de proteínas a una tasa creciente. No obstante, el ratio de producción de esta información estructural es todavía bajo comparado con el relativo a secuencia: en la misma fecha, el número de estructuras de proteínas depositadas en el *Protein Data Bank* (PDB; Berman *et al*. 2007) asciende a 70.107 entradas.

En el plano fenotípico se han llevado a cabo algunos intentos de caracterización fenotípica a escala genómica como los de Winzeler *et al.* (1999; caracterización fenotípica a gran escala de los noqueados de casi un tercio de las pautas abiertas de lectura de *Saccharomyces cerevisiae*), Sönnichsen *et al.* (2005; evaluación fenotípica a gran escala basado en ARNs de interferencia para identificar todos los genes implicados en las primeras dos rondas de división celular en el embrión de *Caenorhabditis elegans*) o más recientemente en Neumann *et al.* (2010; combinación de ARNs de interferencia, microscopía de secuencia temporal y análisis computacional de imágenes para caracterizar fenotípicamente 21.000 genes codificantes de humanos en procesos celulares como la división, la migración o la muerte celular).

Sin embargo, a un nivel mayor de detalle, la caracterización experimental de la función de una proteína y sus correspondientes atributos moleculares sigue siendo difícil de automatizar, conllevando por lo general un elevado coste en tiempo y recursos. Estas limitaciones unidas al enorme número de secuencias y estructuras para las cuales se carece de información funcional detallada, han incentivado -de forma explosiva en los últimos 30 años- el desarrollo de técnicas computacionales que permitan inferir información biológica relevante a partir de los datos disponibles. En esta tesis estamos interesados en particular en el

estudio computacional de los detalles moleculares de las proteínas relacionados con su función bioquímica.

I.2. Sobre el concepto de función de proteínas

La función de las proteínas es un concepto complejo cuyo significado depende del contexto en el que se estudia. Si consideramos p.ej. una proteín-quinasa, la función quinasa desde un punto de vista catalítico corresponde a la fosforilación de un grupo hidroxilo de un determinado sustrato. A nivel fisiológico, esa proteína podría estar participando de una ruta de señalización, interactuando con otras proteínas. La misma proteína expresada en compartimentos celulares o tipos de tejidos distintos podría también considerarse funcionalmente diferente desde el punto de vista de la regulación celular.

El abordaje computacional a gran escala de la función de las proteínas, requiere disponer de recursos en lo que se definan y sistematicen sus diferentes aspectos funcionales. Actualmente, la iniciativa más comprehensiva en este sentido es la del consorcio *Gene Ontology* (GO, Ashburner *et al.* 2000) en el que se ha desarrollado un vocabulario controlado y estructurado para describir la función de cualquier producto génico desglosada en tres niveles interdependientes: a) su acción molecular, como el mecanismo catalítico en el caso de los enzimas; b) el proceso biológico en el que está implicada, el cual requiere generalmente la acción concertada de un conjunto de ellas; y c) los compartimentos celulares (y tejidos en su caso) en los que está presente. En GO, estas categorías se desarrollan con la adición de clases y términos que se relacionan entre sí dando lugar a una estructura similar a un árbol (formalmente, un grafo acíclico dirigido). Sobre este grafo, los términos descriptivos GO se conectan de abajo (un nivel de detalle mayor) a arriba (clases más generales) siguiendo relaciones “de hijo a padre” del tipo “es un…” o “es parte de…", en la que cada “hijo” puede tener diversos “padres”. De este modo, la secuencia anotada con un término GO al nivel más detallado hereda los términos GO de los que ese es hijo y así sucesivamente hasta el nivel más general. En total, GO acumula más de 26.000 términos distribuidos a lo largo de los tres niveles mencionados. De forma importante, GO provee una estimación del grado de confianza atribuido a cada anotación. El principal inconveniente de GO es la dificultad para establecer medidas de parecido funcional entre las anotaciones de dos proteínas debido a que no todos sus términos aportan el mismo nivel de detalle, los hay parcialmente redundantes y existen partes del grafo acíclico más desarrolladas que otras (Alterovitz *et al*. 2010)

En el caso particular de los enzimas (de especial interés para esta tesis) la descripción de su actividad catalítica se ha sistematizado en llamado “código enzimático” (EC) surgido en 1955 del Congreso Internacional de Bioquímica. En este congreso se creó la llamada *Enzyme Commission* para detallar y clasificar

de forma jerárquica mediante 4 dígitos numéricos la descripción de un enzima y los detalles moleculares de su actividad catalítica. El primer dígito de un código EC distingue entre las siguientes seis grandes clases de reacciones: EC 1: Oxidoreductasas; EC 2: Transferasas; EC 3: Hidrolasas; EC 4: Liasas; EC 5: Isomerasas y EC 6: Ligasas. El significado de los siguientes 3 dígitos varía en función de cada una de estas seis clases. Así, el segundo dígito indica el sustrato (en el caso de las oxidoreductasas), el grupo químico que se transfiere (en el caso de las transferasas), el tipo de enlace implicado (en el caso de las hidrolasas, liasas y ligasas) o el tipo de reorganización (en el caso de las isomerasas). P.ej. EC 3.4 denota una enzyma hidrolasa que actúa sobre enlaces peptídicos. El tercer y cuarto dígito aportan progresivamente un nivel mayor de detalle respecto a la reacción catalizada, los sustratos o productos, cofactores, etc.

I.3. Sobre la definición de "residuo funcional"

La función de las proteínas depende normalmente de un subconjunto de sus residuos que, localizados en una región específica de su estructura, interaccionan con factores externos. En los enzimas, por ejemplo, unos pocos residuos situados en el centro activo pueden determinar la reacción catalítica. Este es el caso de las serín-proteasas en las que la tríada de residuos Ser-His-Asp lleva a cabo los pasos clave para la catálisis. En otros casos, el conjunto de residuos funcionales puede constituir regiones de mayor tamaño, como las regiones de interacción proteína-proteína, sitios de unión a pequeños ligandos, a ácidos nucleicos, etc.

De forma general, los residuos funcionales de una proteína pueden definirse como aquellos residuos que determinan de forma directa su mecanismo molecular de acción de tal modo que su rol biológico sea acometido. Esto es, un cambio del tipo de aminoácido en esas posiciones (excepto en ocasiones por algunos aminoácidos compatibles) tendría un efecto potencial en la función de la proteína y, a mayor escala, en el fenotipo del organismo. No obstante, no suelen considerarse como residuos funcionales aquellos cuya influencia en la función es indirecta, por ejemplo en aspectos tan cruciales como su estructura tridimensional. Así, un residuo cuya mutación en el núcleo estructural impida el plegamiento de la proteína y, por tanto, su funcionamiento adecuado, es importante funcionalmente pero no suele incluirse dentro de la definición de residuo funcional.

Al igual que en el caso de las anotaciones funcionales de proteínas, se han desarrollado diferentes repositorios informatizados donde se recoge la anotación funcional de residuos obtenida a partir de información experimental y/o estructural, aunque para una relación exhaustiva en cada caso concreto sigue siendo aconsejable acudir a la literatura científica. En el caso de los enzimas

existen diferentes recursos como las anotaciones recogidas en UniProt (Apweiler et al. 2004), el PDB o de forma más estructurada en PDBsite (Ivanisenko *et al.* 2005). Como repositorio de residuos anotados destaca el *Catalytic Site Atlas* (Porter *et al.* 2004) que recoge de forma supervisada sitios catalíticos asociados a estructuras 3D derivados del PDB y de la literatura. En el caso de los residuos implicados en la unión a pequeños ligandos, la mayor fuente de anotación de residuos funcionales se deriva de los contactos atómicos observados en las estructuras del PDB, a partir de los cuales se han desarrollado diversas bases de datos con diferentes niveles de organización, p.ej. Ligand-depot (Feng *et al.* 2004; orientada a ligandos, aporta características químicas y geométicas pero sin relacionarlos con las estructuras a las que unen), Relibase (Hendlich *et al.* 2003; de los recursos más completos para explorar complejos proteína-ligando aunque alguna de sus funcionalidades no son de libre acceso), Ligbase (Stuart *et al.* 2002; recoge patrones tridimensionales de residuos en una conformación espacial determinada observados en sitios funcionales previamente conocidos). De especial interés para los análisis que se hacen en esta tesis es la base da datos FireDB (López *et al.* 2007), desarrollada en nuestro grupo, que reúne residuos funcionales de estructuras de proteínas tanto catalíticos (extraídos del *Catalytic Site Atlas*) como de unión a pequeños ligandos (derivados del PDB). De forma importante, FireDB es capaz de distinguir los sitios de unión biológicamente relevantes filtrando aquellos que representan artefactos producidos durante el proceso de cristalización. Considerando estas características, FireDB es la colección más completa y fiable de este tipo.

I.4. Importancia funcional de las interacciones proteína-proteína

Junto a los aspectos funcionales relacionados con la actividad catalítica y la unión a pequeños ligando, en esta tesis estamos interesados en los aspectos funcionales de las proteínas que implican la interacción física con otras proteínas. Las interacciones proteína-proteína intervienen en el control del ciclo celular, la diferenciación celular, el plegamiento de proteínas, la señalización y el transporte, así como en la transcripción, traducción y modificaciones post-traduccionales. Pueden además alterar las propiedades catalíticas de los enzimas, permitir la canalización de secuencias de reacciones, habilitar nuevos sitios activos, inactivar o destruir una proteína, cambiar su especificidad, tener un papel regulatorio, etc.

En los últimos años se han llevado a cabo importantes esfuerzos para determinar el conjunto de las interacciones proteína-proteína que se producen en las células de diferentes organismos, tanto a través de experimentos a pequeña escala (mediante experimentos individuales diseñados específicamente para la identificación de un pequeño número de interacciones concretas; Golemis 2002) como mediante técnicas masivas como el sistema de doble híbrido en levadura (*yeast two-hybrid*, Fields y Song 1989; Uetz *et al.* 2000; Ito *et al.* 2001) o la co-

precipitación e identificación de complejos con espectrometría de masas (p.ej. Gavin *et al.* 2002; Ho *et al.* 2002). En paralelo se han desarrollado diferentes bases de datos que tratan de aglutinar y sistematizar, para numerosos organismos, las diferentes evidencias experimentales de la interacción física entre proteínas, p.ej. DIP (Xenarios et al. 2002), MINT (Chatraryamontri et al. 2007) IntAct (Aranda et al. 2010), MIPS (Mewes et al. 1997), BIOGRID (Stark et al., 2006), etc. (véase Klingstöm y Plewczynski 2010 y Lee et al. 2007, para una relación más exhaustiva). BIOGRID destaca por ser el repositorio general de interacciones más amplio, recogiendo evidencias experimentales de cualquier tipo, tanto a pequeña como a gran escala, para numerosos organismos, aunque a costa de incluir información con un nivel menor de detalle. Entre las que realizan un proceso de supervisión manual realizado por expertos destacan DIP, MINT e IntAct. DIP combina diferentes fuentes de información para derivar un conjunto de interacciones consistente en cada uno de los organismos que recoge. Para ello realiza dos niveles de supervisión, uno automático y otro supervisado manualmente, identificando un subconjunto de interacciones de alta confianza (DIP core). IntAct (Aranda *et al.* 2010) combina información de la literatura y de conjuntos de interacciones experimentales para diferentes organismos y MINT (Chatraryamontri et al. 2007) extrae las interacciones de proteínas a partir de la información publicada en la literatura, con especial énfasis en mamíferos. Si bien la supervisión manual de estas iniciativas incrementa notablemente la calidad de las anotaciones (eliminando muchos falsos positivos derivados de las técnicas a gran escala), no está libre de error a tenor de las disparidades encontradas entre ellas (Cusick et al. 2009).

Desde el punto de vista estructural, el PDB es la mayor fuente de información de interacciones proteína-proteína permitiendo derivar de las estructuras de los complejos depositados los residuos implicados en la interfaz de unión. Diferentes bases de datos recogen estos residuos, tomando como referencia proteínas completas (Pibase; Davis y Sali, 2005) o dominios (3did , Stein *et al*. 2005; iPfam; Finn *et al.* 2005). De forma importante, iniciativas como PiQSi (Levy *et al.* 2007) supervisan manualmente los complejos depositados en el PDB tratando de verificar que respondan a estructuras cuaternarias biológicamente relevantes y no a artefactos producidos durante el proceso de cristalización.

Por otra parte, a partir de los resultados obtenidos mediante estudios termodinámicos consistentes en mutaciones a Alanina de los residuos de la interfaz proteína-proteína (*alanine scans*), se ha postulado que sólo un pequeño subconjunto de los residuos son esenciales para el reconocimiento y unión entre interactores (Clackson y Wells, 1995; Bogan y Thorn, 1998; Thorn y Bogan, 2001). Este subconjunto crítico de residuos (identificados generalmente por contribuir con más 2 kcal/mol a la interacción) tienden a estar agrupados (Moreira *et al.*, 2007) por lo que ha recibido el nombre común de *hot spots*

(puntos calientes). La caracterización y significación biológica de este tipo de regiones es todavía ambigua, por una parte por la escasez de los datos experimentales (véanse las bases de datos ASEdb, Thorn y Borgan 2001, y BID, Fischer et al. 2003) y por otra por los inconvenientes intrínsecos de los experimentos de mutación a Alanina. Así, se ha señalado que las mutaciones simples no siempre aportan información exacta respecto a la contribución de un residuo a la interacción (Vaughan *et al.*, 1999; Reichmann *et al.*, 2005); que las propiedades fisicoquímicas de las proteínas no equivalen a la suma de las aportaciones de sus residuos individuales (Horovitz 1996) y que los *hot spots* no siempre forman parte de la interfaz sino que su acción puede ser indirecta reorientado las cadenas laterales de otros residuos (DeLano 2002).

I.5. Del concepto de "familia de proteínas homólogas"

Desde un punto de vista evolutivo, se dice que dos o más proteínas son homólogas si comparten un ancestro común (Fitch 1970). Las proteínas homólogas, a lo largo de la evolución divergen en secuencia como resultado de una variedad de eventos genómicos (incluyendo mutaciones puntuales, duplicaciones, deleciones, etc.). La variabilidad genética resultado de estos eventos, cuando afecta a genes que codifican proteínas, sufre la presión selectiva asociada a la estructura tridimensional y función de dichas proteínas en contextos celulares específicos (Ohno 1970; Fitch 1970; Rossmann y Argos 1981; Blake 1983; McCarthy y Hardie 1984; Gilbert 1985; Tatusov *et al.* 1997). Si bien durante este proceso de divergencia la estructura tridimensional tiende a mantener una arquitectura similar (Chothia y Lesk 1986), frecuentemente la divergencia en secuencia entre proteínas homólogas refleja la realización de funciones diferentes (Henikoff *et al.* 1997). En esta tesis estamos interesados en la comprensión de los aspectos funcionales que han determinado esa divergencia en secuencia, cómo esta se traduce en la adquisición de funciones específicas que diferencian a los homólogos entre sí y cuáles son los residuos que las determinan.

La homología entre proteínas se infiere mediante criterios heterogéneos basados en la práctica en similitud (global o sólo en determinados rasgos) en secuencia y/o estructura , para lo cual se hace uso de alineamientos apareados o múltiples. En ellos, los aminoácidos de las proteínas implicadas son dispuestos de forma tal que pueden considerarse equivalentes bajo una perspectiva evolutiva (corresponden a una misma posición en la secuencia ancestral) y/o estructural (equivalencia espacial entre posiciones de proteínas diferentes).

Se han utilizado diferentes aproximaciones para establecer conjuntos de proteínas homólogas dando lugar a numerosas bases de datos de familias de proteínas. Estas pueden ser el resultando de búsquedas en secuencia basadas en diferentes metodologías como el uso de expresiones regulares (PROSITE,

Sigrist *et al.* 2002), perfiles (PRINTS, Attwood *et al*. 2003) o Modelos Ocultos de Markov (HMMs; Eddy 1996) para dominios de proteínas (Pfam, Bateman *et al*. 2004; SMART, Schultz *et al*. 1998) o para secuencias completas (TIGRFAM, Haft *et al*. 2007; PIRSF, Wu *et al*. 2004; PANTHER, Thomas *et al*. 2003). También se han creado clasificaciones de proteínas basadas en semejanza estructural sobre las estructuras del PDB, como SCOP (Andreeva *et al*. 2004) o CATH (Orengo *et al*. 1997) a partir de las cuales se han generado nuevas familias de proteínas mediante búsquedas en secuencia con HMMs (SUPERFAMILY, Wilson *et al*. 2007, derivada de SCOP; Gene3D, Yeats *et al*. 2006, derivada de CATH). Cabe destacar aquí como recurso de familias de proteínas la base de datos InterPro (Mulder *et al*. 2007), que ha resultado de la integración y sistematización de muchos de los recursos provistos por las bases de datos de familias de proteínas anteriores.

El alineamiento múltiple de las secuencias (MSA) de una familia de proteínas permite estudiar la divergencia acumulada en una familia de proteínas a lo largo de la evolución. Estos alineamientos son importantes fuentes de información funcional y estructural (Devos *et al.* 2002; Valencia 2005) puesto que muestran los cambios permitidos o las restricciones mantenidas por la evolución en cada posición equivalente. Los alineamientos múltiples permiten –además de observar la divergencia total considerando todos los residuos de las proteínas- identificar posiciones con una señal de restricción evolutiva. El primer tipo de patrones estudiados fueron las posiciones mayoritariamente conservadas. Este tipo de posiciones se interpretan como residuos importantes para la estructura y/o función de la proteína puesto que no se han dado cambios en ella a lo largo del proceso evolutivo. Estas posiciones fueron los primeros indicadores de funcionalidad (Zuckerkandl y Pauling, 1965) y se relacionan con todo tipo de sitios funcionales: sitios catalíticos (Zvelebil *et al.* 1987), de unión a ligando (Ouzounis *et al.* 1998), de interacción proteína-proteína (Valdar y Thornton 2001), de unión a ácidos nucleicos (Aloy *et al.* 2001), etc. Sin embargo, no todas las posiciones conservadas están relacionadas con función sino que muchas lo están por requerimientos estructurales constituyendo el núcleo estructural de la proteína (Chothia y Lesk 1986; Sander y Schneider 1991; Schueler-Furman y Baker 2003, Chelliah *et al.* 2004). En general, se asume que las posiciones conservadas a lo largo de las proteínas homólogas de una familia están implicadas en las características funcionales y/o estructurales comunes a todas ellas, no pudiendo explicar en cambio -por sí solas- las diferentes especificidades funcionales entre proteínas homólogas comentadas anteriormente.

Se han desarrollado numerosas metodologías para evaluar el grado de conservación de una posición entre proteínas homólogas (Valdar 2002) desde el más simple porcentaje de identidad del aminoácido mayoritario en la columna, la conservación teniendo en cuenta propiedades fisicoquímicas o matrices de

sustitución (p.ej. una posición con Argininas, Lisinas e Histidinas) a métricas más elaboradas: medidas de entropía como la de Shanon o la relativa respecto a una distribución de fondo (Capra y Singh 2007), la varianza respecto a la distribución media de aminoácidos en el alineamiento completo (Pei y Grishin 2001) u otros métodos que consideran la filogenia de las proteínas a fin de evitar resultados sesgados por una distribución desigual de secuencias en el alineamiento (p.ej. presencia de muchas secuencias muy similares; Pupko *et al.* 2002; Engelen *et al.* 2009).

I.6. Divergencia en secuencia entre proteínas homólogas y su relación con la divergencia funcional

Por lo general, las familias de proteínas homólogas comparten cierto grado de características funcionales generales, diferenciándose en otros aspectos más específicos. Diferentes estudios han tratado de cuantificar la similitud/diversidad funcional entre dos proteínas homólogas en función del porcentaje de identidad en secuencia que presentan.

En un estudio a gran escala, Devos y Valencia (2000) determinaron el porcentaje promedio de identidad de secuencia (Seq ID) a partir del cual dos proteínas comparten i) los dos primeros dígitos EC (15% Seq ID); ii) la mitad de palabras clave descriptivas de función extraídas de SwissProt (Apweiler *et al*. 2004) (20% Seq ID); iii) el 50% de los aminoácidos en las posiciones que conforman el sitio de unión a ligando (30% Seq ID); y iv) una probabilidad media del 70% de compartir la misma función celular, según las clases propuestas por Riley (1993) (entre 30-70% Seq ID).

Todd *et al.* (2001), combinando información en secuencia y estructura para definir grupos de proteínas homólogas ("superfamilias" en la definición de CATH, Green *et al*. 2007) encontraron que en la mayoría de ellos se encuentra variación en la función enzimática, principalmente en lo relativo a la especificidad de sustrato, si bien el tipo de reacción química se conserva por lo general. En este mismo estudio, se determinó que la variación en el código EC entre pares de proteínas homólogas es infrecuente por encima del 40% de identidad de secuencia y que, por encima del 30%, sus tres primeros dígitos pueden predecirse con una exactitud del 90%. Por debajo de ese límite, en cambio, la diversidad funcional es significativa y es necesario recurrir a información estructural para entender el detalle molecular de esas diferencias. Posteriormente, Tian y Skolnich (2002) corrigieron a la baja estos resultados ajustando el sesgo producido por la sobrerrepresentación de algunas familias y establecieron que para transferir con cierta garantía los tres primeros dígitos hace falta al menos el 40% de identidad de secuencia. Cuando se agrupan enzimas que comparten sus tres primeros dígitos, estos mismos autores estimaron que por encima del 60% de identidad de secuencia se comparten los

4 dígitos en alrededor del 90% de los casos. Por otra parte -de forma importante para esta tesis- las posiciones conservadas en las familias de proteína tienden a formar parte del sitio de unión de pequeños ligandos y de sitios catalíticos (Zvelebil *et al.* 1987; Ouzounis *et al.* 1998), de forma consistente para un gran número de familias e independientemente de la formas de cálculo de la conservación (Capra y Singh, 2007; Fischer *et al.* 2008; Manning *et al.* 2008).

Sangar *et al*. (2007) estudiaron también la relación entre divergencia en secuencia y función, si bien en aspectos de esta última más comprehensivos a partir de la descripción de términos funcionales de *Gene Ontology* comentada anteriormente. Los autores encontraron que, por encima del 50% de identidad de secuencia, la divergencia funcional decae rápidamente, siendo raros los casos con anotaciones funcionales totalmente diferentes.

No obstante la relación general entre divergencia en secuencia y divergencia funcional, cabe mencionar que existen casos en los extremos de estas distribuciones en los que proteínas muy parecidas presentan grandes diferencias funcionales y, al contrario, proteínas muy distantes que conservan su función.

I.7. Interacciones proteína-proteína comunes y específicas entre proteínas homólogas

Entre los diferentes aspectos funcionales que pueden determinar la divergencia en secuencia en una familia de proteínas, para esta tesis son particularmente relevantes los relacionados con la interacción física entre proteínas. Diversos estudios han tratado de evaluar el grado de conservación de la interacción con terceras proteínas que tienen dos proteínas homólogas, bien intraespecie o entre especies distintas en función de su divergencia en secuencia:

Mika y Rost (2006) mostraron que las interacciones entre pares de proteínas de un organismo tienden a conservarse en otras especies cuando la similitud en secuencia con sus correspondientes homólogos es muy elevada. En ese mismo estudio, los autores estimaron que la inferencia de pares de interactores a través de la similitud en secuencia es significativamente más fiable para pares de proteínas dentro del mismo organismo que entre organismos diferentes. Los resultados de Mika y Rost (2006) son consistentes con el elevado ratio de pérdidas y ganancias de interacciones proteína-proteína (re-cableado) observado al comparar las redes de interacción de especies distintas (Wagner 2001; Wagner 2003; Suthram *et al.*, 2005; Beltrao y Serrano 2007; Shou et al. 2011).

En lo relativo a los módulos funcionales conformados por la interacción de conjuntos de proteínas en forma de complejos estables, si bien se puede encontrar una variación significativa entre especies (Snel y Huynen 2004), se ha visto que están más conservados cuanto mayor sea su interconexión (Wutchy *et al.* 2003). También en un mismo organismo se puede encontrar una cantidad considerable de módulos funcionales duplicados que comparten una función general común aunque con diferentes especificidades funcionales (Pereira-Leal y Teichmann, 2005). Estudios más recientes apuntan a la pérdida o adquisición genómica de proteínas para explicar la diferente composición de los módulos funcionales entre especies, sugiriendo sin embargo un alto porcentaje de conservación en módulos funcionales equivalentes cuando los diferentes organismos han mantenido las proteínas implicadas (van Dam y Snel 2008).

Por otra parte, a partir de información estructural, Aloy *et al.* (2003) observaron que los interólogos (los pares de homólogos que sí interaccionan en diferentes organismos) tienden a interaccionar en topologías equivalentes cuando su similitud en secuencia es mayor del 30%, si bien los contactos equivalentes residuo-residuo se comparten sólo parcialmente (Aloy y Russell 2002). También en el caso de los homo-oligómeros, los estudios a gran escala de estas estructuras han mostrado que su estructura cuaternaria (su forma de simetría) se conserva en el 70% de aquellos casos cuyo nivel de similitud en secuencia es mayor al 30% (Levy *et al.*, 2008).

Si las interfaces proteína-proteína están o no más conservadas que el resto de la superficie de forma generalizada (esto es, para un número significativo de casos) ha sido una cuestión recurrente a lo largo de los últimos 15 años sin haberse alcanzado todavía un amplio consenso al respecto. Los primeros análisis, realizados sobre conjuntos pequeños de interfaces, arrojaron resultados contradictorios tanto en positivo (Jones y Thornton, 1996; Jones y Thornton, 1997; Valdar y Thornton 2001) como en negativo (Grishin y Phillips, 1994; Caffrey et al., 2004). Los estudios se han ido beneficiando del número de estructuras cristalizadas y los últimos publicados (Choi *et al.* 2009; Engelen *et al.* 2009; Guharoy y Chakrabarti 2010) -ya a mayor escala- constatan un grado significativo de conservación mayor en las interfaces proteína-proteína, tanto homodiméricas como heterodiméricas, que en el resto de la superficie proteica, si bien puede observarse mucha heterogeneidad entre familias distintas (Panjkovich y Aloy 2010). La tendencia de las posiciones conservadas a formar parte de ambos tipos de interfaz (homodiméricas y heterodiméricas) es además robusta a las diferentes metodologías utilizadas para el cálculo de la conservación comentadas anteriormente (Manning *et al.* 2008).

No obstante, las conclusiones adquieren matices en función de cómo se defina la interfaz. Así, la conservación es en general de mayor grado en su núcleo central (residuos totalmente cubiertos en la unión) que en el anillo exterior

(residuos parcialmente cubiertos en la periferia) (Caffrey *et al*. 2004; Bordner y Abagyan 2005; Guharoy y Chakrabarti 2005; Guharoy y Chakrabarti 2010). La señal de conservación es también más clara cuando se tienen en cuenta las diferentes interfaces que una proteína dada puede tener con interactores diferentes (Choi *et al*. 2009).

Por otra parte, la tendencia de los *hot spots* a estar agrupados en superficie coincide con la observación de que los residuos conservados en la interfaz tienden a estar espacialmente agrupados (Jones y Thornton 1997) tanto en interacciones homodiméricas como heterodiméricas (Guharoy y Chakrabarti 2010; Hwang *et al*. 2008). Esto ha llevado a indagar el grado de solape entre ambas agrupaciones de residuos. Si bien algunos trabajos apuntan a un solape importante entre hot spots y agrupaciones de residuos conservados (Li *et al.* 2004; Halperin et al. 2004; Keskin *et al.* 2005), otros más recientes y relativamente a mayor escala muestran una correlación mucho más modesta (Guharoy y Chakrabarti 2010; Ofran y Rost 2007).

I.8. Las mutaciones correlacionadas y su relación con las interacciones entre proteínas

Las denominadas "mutaciones correlacionadas" son otro tipo de información relacionada con funcionalidad que puede extraerse de los alineamientos múltiples de proteínas. Este tipo de posiciones son conceptualmente importantes en esta tesis y con ellas se pondrán en relación algunos aspectos que se presentan.

Las mutaciones correlacionadas en un MSA se definen como pares de posiciones en que muestran una tendencia a mutar de forma coordinada (Pazos *et al.* 2007). Los patrones coordinados en estas posiciones reflejarían cambios compensatorios desencadenados por mutaciones desfavorables generadas por deriva genética o relacionados con su adaptación a nuevas funciones. Estas posiciones permiten observar importantes aspectos de la co-evolución en proteínas (Pazos y Valencia 2008). Se pueden distinguir dos tipos de mutaciones correlacionadas: las detectables dentro de una misma familia de proteínas (a partir de su MSA) y las que se producen entre dos familias de proteínas (a partir de la comparación entre los dos MSAs correspondientes). Para su identificación se han utilizado diferentes medidas de covariación incluyendo coeficientes de correlación, información mutua y el cálculo de la desviación entre las distribuciones marginales y condicionales de los tipos de aminoácidos en las columnas del MSA (véase Halperin *et al.* 2006 para una descripción detallada de las diferentes métricas).

Las posiciones correlacionadas intra-familia muestran una tendencia –si bien moderada- a estar próximas en estructura (Göbel *et al.*, 1994; Olmea y

Valencia, 1997) y han sido utilizadas en combinación con las posiciones conservadas para el reconocimiento y la predicción de plegamientos estructurales (Olmea *et al.* 1999; Larson *et al.* 2000; Fariselli *et al.* 2001). No obstante, estudios posteriores han mostrado que, por sí solas, su capacidad para predecir contactos intra-proteína apenas excede el 20% (Fodor y Aldrich 2004). Además de entre contactos directos, también se han encontrado mutaciones correlacionadas entre posiciones distantes y funcionalmente importantes dentro de una misma proteína (Fares y Travers 2006), las cuales se han relacionado en ocasiones con acoplamientos energéticos implicados en la propagación de señales alostéricas (Lockless y Ranganathan 1999; Kass y Horovitz 2002; Shi *et al.* 2006). Por su parte, Kowarsch *et al.* 2010 han mostrado que, cuando ocurre una mutación en posiciones con un patrón de cambio correlacionado, la probabilidad de que sea causante de enfermedad está por encima del azar.

En el caso de los enzimas, Marino-Buslje *et al.* (2010) han mostrado a gran escala que los residuos cercanos a los sitios catalíticos muestran una sobrerrepresentación de posiciones con un patrones de cambio correlacionado, apuntando a su importancia en el mantenimiento del entorno estructural del sitio activo. En el caso de pares de proteínas que interaccionan físicamente, Pazos *et al.* (1997) observaron que sus interfaces de unión acumulan posiciones correlacionadas inter-familia acopladas espacialmente, una tendencia que trabajos más recientes confirman a mayor escala (Yeang y Haussler 2007; Madaoui y Guerois 2008). Por otra parte, el análisis de la co-evolución en secuencia entre familias permite detectar interacciones físicas proteína-proteína (Pazos y Valencia 2001; Pazos *et al.* 2005; Juan *et al.* 2008), predecir pares de interacciones específicas entre las proteínas de dos familias (Ramani y Markote 2003; Izarzugaza *et al.* 2006) así como residuos en ambas proteínas interactoras en contacto directo (Weigt *et al.* 2009)

I.9. Subfamilias de proteínas y posiciones determinantes de especificidad funcional (SPDs)

El concepto de conservación se amplió posteriormente para considerar posiciones con un patrón de conservación diferencial entre distintas sub-agrupaciones de proteínas homólogas, esto es: posiciones conservadas dentro de cada grupo pero con un tipo de aminoácido diferente entre los diferentes grupos de una misma familia (**Fig. 1**). El que estas posiciones estén conservadas sugiere una importancia funcional, mientras que el hecho de que el tipo de aminoácido sea diferente indica que esta importancia es específica de cada grupo en relación al criterio seguido para definirlos.

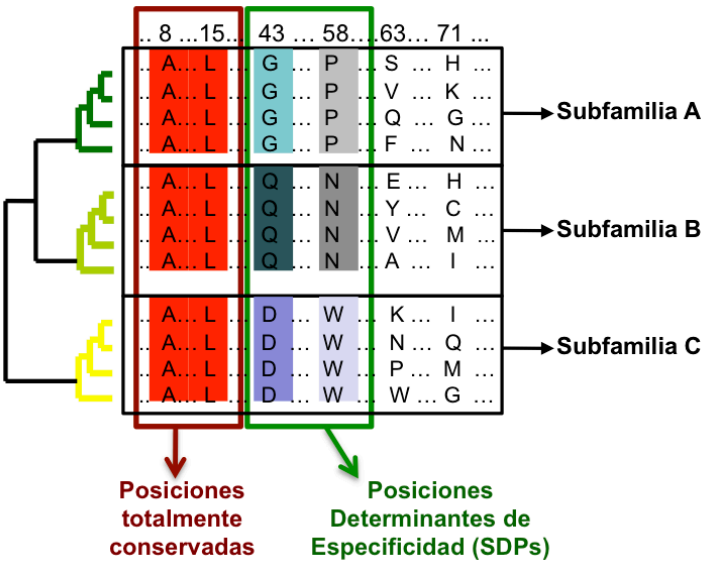


Fig. 1. Representación esquemática de un alineamiento múltiple de proteínas. La figura muestra un alineamiento de proteínas cuyas relaciones de parecido en secuencia están representadas por el árbol a la izquierda y se traducen en una organización en 3 subfamilias. Pueden observarse posiciones totalmente conservadas y posiciones con un patrón de conservación diferencial que se ajusta a las subfamilias (SDPs; ver texto)

La definición de los grupos se puede realizar de forma supervisada -en base a clasificaciones funcionales conocidas a priori- o no supervisada. En el caso no supervisado (el más común) se persigue identificar conjuntos de proteínas más similares dentro de grupo que entre grupos. Los parecidos relativos entre proteínas homólogas pueden derivarse de su similitud en secuencia de acuerdo al alineamiento múltiple, o bien a través de la estructura jerárquica ofrecida por el árbol filogenético de la familia. La granularidad de estas agrupaciones se puede explorar de forma gradual en un mismo análisis (considerando particiones sucesivas: desde las de menor número de grupos y mayor número de secuencias por grupo hasta las de mayor número de grupos) o bien establecerse una única partición que refleje de forma óptima el nivel de organización intra-familia, en el caso común que su divergencia en secuencia no sea uniforme sino estructurada (Dayhoff et al. 1983; Murzin *et al.* 1995). Los grupos de proteínas de esa partición óptima son considerados como las subfamilias en las que la familia de proteínas está organizada.

En los apartados anteriores se vio que las familias de proteínas homólogas comparten ciertos aspectos funcionales y presentan también funciones específicas que las diferencian (Henikoff et al. 1997). Se vio también que el grado de conservación de una posición es indicativo de su importancia funcional y que la conservación de las regiones funcionales no es total sino que deja margen a la existencia de posiciones con diferentes tipos de aminoácidos

que puedan asociarse a sus características específicas. Por otro lado, salvo en casos poco frecuentes de convergencia evolutiva, se asume que la adquisición de peculiaridades funcionales distintivas entre proteínas homólogas se refleja en la divergencia en secuencia acumulada en la familia (Devos y Valencia 2000; Rost 2002; Sjölander 2004). A partir de las consideraciones anteriores (y de forma importante para esta tesis), la hipótesis de trabajo generalmente aceptada es que, por una parte, las subfamilias deberían corresponder a grupos de proteínas con especificidades funcionales distintivas y, por otra, que las posiciones con un patrón de conservación diferencial (asociado a la correspondiente organización en subfamilias) serían las posiciones responsables de sus funciones específicas. Por esta razón, a este tipo de posiciones se les ha denominado "posiciones determinantes de especificidad funcional" (SDPs por sus siglas en inglés: *Specificity Determining Positions*; Mirny y Gelfand 2002; Kalinina et al 2004; Donald and Shakhnovich 2005; Pei *et al.* 2006; Capra y Singh 2008)[1].

Los diferentes abordajes metodológicos que se han utilizado para la detección tanto de subfamilias como de SDPs son relevantes para el trabajo que aquí se presenta y en el que específicamente se desarrollan nuevas metodologías con la misma finalidad. Se revisan a continuación estas metodologías para posteriormente analizar el grado de confirmación de la hipótesis por la cual subfamilias y SDPs están asociados a características funcionales específicas.

I.10. Abordajes computacionales para la detección de subfamilias y SDPs en una familia de proteínas homólogas

Se han desarrollado numerosas métodos para la detección de subfamilias y/o SDPs en una familia de proteínas homólogas. Por una lado, existen métodos especializados únicamente en la definición de subfamilias (sin detectar SDPs) los cuales recaen en dos categorías principales: i) los que definen agrupaciones a partir de la similitud en secuencia, p.ej. Ncut (Abascal y Valencia 2002), CD-Hit (Li y Godzik 2006) o el algoritmo de partición de grafos de Donald y Shakhnovich (2005); y ii) los que establecen un corte óptimo sobre el árbol filogenético u otro tipo de agrupación jerárquica de las proteínas de la familia, p.ej. Secator (Wicker et al. 2001), SCI-PHY (Brown et al. 2007) y GEMMA (Lee et al. 2010).

Por otro lado, toda una serie de metodologías están orientadas a la detección de SDPs, algunas de las cuales ofrecen además una partición óptima

[1] Otras denominaciones sinónimas en la literatura en inglés incluyen las siguientes: Tree-determinants (Andrade *et al.* 1997; Bauer et al. 1999; del Sol *et al*. 2003; Tress *et al.* 2005; Carro *et al*. 2006); *specificity determining residues* (SDRs; Mirny y Gelfand 2002; Donald y Shakhnovich 2005; Ye *et al.* 2008); *function-discriminating residues* (FDRs; Lee *et al.* 2007); subfamily-specific sites (Pirovano et al. 2006).

en subfamilias. La mayoría de estos métodos pueden clasificarse en tres grandes abordajes siguiendo a del Sol *et al.* (2003) que se resumen a continuación (en la **Tabla 1** se recogen los principales métodos correspondientes a cada una de ellas junto con un cuarto grupo de métodos que no encajan bien en ninguna de las anteriores; en la **Tabla 2** se resume cómo opera cada uno de los métodos recogidos en la **Tabla 1)**:

i) Métodos que utilizan de forma explícita el árbol filogenético asociado al alineamiento múltiple. Sobre él van rastreando sucesivos cortes a distintos niveles de profundidad, desde la raíz hasta las hojas, generando progresivamente un mayor número de subfamilias con un menor número de secuencias en cada una de ellas. La importancia de una conservación diferencial para una posición dada se evalúa en función del nivel de profundidad en el árbol al que se han definido los grupos.

ii) Métodos basados en el principio por el cual las regiones funcionalmente importantes preservan (o reproducen) las distancias relativas entre las proteínas de la familia calculadas sobre su secuencia completa (Elcock 2001; Livesay *et al.* 2003; del Sol *et al*. 2003). En estos métodos, las posiciones que mejor explican la separación observada entre las proteínas de la familia son consideradas como responsables de sus diferencias funcionales. En cuanto a su implementación metodológica, este tipo de métodos tratan de localizar posiciones en el MSA cuyas relaciones de semejanza/diferencia entre los tipos de aminoácidos que presentan maximicen su correlación con las semejanzas/diferencias entre el total de posiciones de las secuencias del alineamiento.

iii) El abordaje multivariante de *Sequence Space* (Casari *et al.* 1995). *Sequence Space* fue el primer método no supervisado de detección de residuos funcionales conservados a nivel de subfamilia. Este elegante abordaje parte de una representación vectorial del MSA en dos espacios de alta dimensión, uno de secuencias y otro de residuos. A continuación, se transforman estos espacios mediante un tratamiento con Análisis de Componentes Principales (PCA) obteniendo espacios equivalentes de dimensión reducida íntimamente relacionados. Estos espacios permiten observar las principales fuentes de variabilidad en el MSA. En el espacio de secuencias, proteínas con parecido de secuencia elevado tienden a agruparse en la mismas regiones, permitiendo identificar la organización interna en subfamilias. Simultáneamente, los residuos responsables de la separación en subfamilias (aquellos con un patrón de conservación diferencial) se localizan en las regiones equivalentes en el espacio de residuos.

Categoría metodológica	Nombre del método		Referencia	Detección de Subfamilias
Uso explícito de árboles filogenéticos	Familia de Métodos Evolutionary Trace (ET)	ET original	Lichtarge *et al.* 1996b	No
		ET modificado	Mihalek *et al.* 2004	No
		ConSurf	Armon *et al.* 2001	No
		Rate4Site	Pupko et al. 2002	No
		Bayesian- Rate4Site	Mayrose *et al.* 2004	No
		ConSeq	Bezerin *et al.* 2004	No
	S-Method		del Sol *et al.* 2003	Si
	SDPsite		Kalinina *et al.* 2009	Si
Correlación de las posiciones con la variabilidad global	MB		del Sol *et al.* 2003	No
	MINER		La *et al.* 2005	No
	3D Cluster Analysis		Landgraf *et al.* 2001	No
	Strong Motif Search Algorithm		Bickel *et al.* 2002	No
Análisis multivariante	Sequence Space		Casari *et al.* 1995	Si
	SS-Method		del Sol *et al.* 2003	No
Otros	Algoritmo SOM modificado		Andrade *et al.* 1997	Si
	K-PAX		Marttinen *et al.* 2006	Si
	SPEL		Pei *et al.* 2006	No
	CEO (Combinatorial Entropy Optimization)		Reva *et al.* 2007	Si

Tabla I.1. Relación de los principales métodos no supervisados para la detección de SDPs a partir de MSAs agrupados por categorías metodológicas (para una descripción de los métodos véase la Tabla I.2)

Tabla I.2. Descripción de los principales métodos no supervisados para la detección de SDPs a partir de MSAs

Métodos que utilizan de forma explícita el árbol filogenético asociado al alineamiento múltiple		
Familia de Métodos Evolutionary Trace (ET) *	**ET original**	ET proporciona un ranking de todas las posiciones del MSA en función del nivel del árbol en el que aparecen conservadas diferencialmente en todos los grupos de esa partición. Con los mejores rankings aparecen las posiciones totalmente conservadas y en los peores las posiciones totalmente variables (conservadas diferencialmente en una partición en la que hay tantos grupos como secuencias en el alineamiento). En su implementación original, ET representa el grado de conservación de forma cualitativa.
	ET modificado	A partir del método ET original, incorpora una medida de entropía como indicador cuantitativo de conservación a la hora de puntuar y ordenar las posiciones del alineamiento.
	ConSurf	A partir del método ET original, incorpora mejoras la construcción del árbol filogenético y tiene en cuenta la similitud fisicoquímica de los aminoácidos en la evaluación de la conservación de las posiciones.
	Rate4Site	A partir de ConSurf, introduce mejoras para corregir los sesgos derivados de una distribución desigual de secuencias en el MSA y pesa la conservación de las posiciones según su ratio de evolución.
	Bayesian-Rate4Site	A partir de Rate4Site, el cálculo de los ratios de evolución de las posiciones pasa de calcularse mediante máxima verosimilitud a utilizar técnicas bayesianas.
	ConSeq	Similar a Bayesian-Rate4Site pero, en lugar de utilizar información estructural para identificar conglomerados de posiciones conservadas, se utiliza un predictor de accesibilidad al solvente basado en secuencia (Fariselli y Casadio, 2001).
	Otros:	Aloy et al. 2001, Nimrod et al. 2005, Sankararaman y Sjölander 2008.
* La familia de métodos Evolutionary Trace acepta todas las particiones del árbol, desde la raíz (todas las secuencias en una familia) hasta las hojas. En consecuencia, sus predicciones contemplan un nivel de conservación decreciente de forma gradual, donde las posiciones mejor puntuadas son las totalmente conservadas. Esta característica sitúa a los métodos de esta familia en un terreno intermedio entre aquellos que detectan posiciones conservadas y aquellos que detectan SDPs.		
S-Method	Este método explora sucesivas particiones del árbol filogenético asociado a un MSA desde la inmediata posterior a la raíz hasta las hojas, de forma similar a la de Evolutionary Trace, pero aquí se selecciona una sola partición. Los SDPs se definen como aquellas posiciones cuyos aminoácidos muestren al menos un 85% de identidad dentro de subfamilia y sean diferentes entre subfamilias. La partición en el árbol que se selecciona es aquella que minimiza la dependencia entre el número arrojado de SDPs y el número de posiciones conservadas dentro de subfamilia (este último tiende a aumentar de forma trivial a medida que se desciende hacia las hojas). Para ello se maximiza la entropía relativa entre ambas distribuciones.	
SDPsite	Explora diferentes particiones derivadas del árbol filogenético asociado al MSA. En cada partición, SDPpred utiliza información mutua (*mutual information*) para evaluar el ajuste de los tipos de aminoácidos presentes en una posición con una partición en subfamilias determinada A partir de una distribución de fondo generada mediante muestreos aleatorios de posiciones del MSA, se definen los SDPs de la partición como el conjunto de posiciones mejor puntuadas con menor probabilidad de serlo por azar. La comparación de las probabilidades asociadas al conjunto de SDPs de las distintas particiones a diferentes niveles del árbol permite seleccionar la solución definitiva	

Métodos que analizan la correlación de las posiciones con la variabilidad global	
MB	Se codifica el patrón de cambio de una posición (su llamado "comportamiento mutacional") en forma de matriz de semejanza a partir de alguna medida sobre el parecido entre cada par de aminoácidos. Análogamente, el patrón de divergencia entre las secuencias completas se representa en otra matriz de semejanza codificando ésta el parecido global entre cada par de proteínas del MSA. Ambas matrices de semejanza son entonces comparadas a través de alguna medida de correlación (p.ej. de Spearman o de Pearson) asignando una puntuación a cada posición del MSA. Las posiciones con mayor puntuación se predicen como las responsables del patrón de diversidad funcional presente en el MSA.
MINER	Búsqueda de ventanas en secuencia del MSA (alrededor de 5 posiciones) que reproducen la topología del árbol filogenético de la familia.
3D Cluster Analysis	Este método requiere de información estructural además del MSA. Se buscan conjuntos de posiciones agrupadas en superficie cuyas relaciones de semejanza correlacionen con los parecidos globales entre las secuencias del alineamiento.
Strong Motif Search Algorithm	Se rastrean todos los subconjuntos posibles que se pueden formar con las secuencias del alineamiento, identificando conjuntos de posiciones conservadas dentro de ellos con un tipo de aminoácido diferente a las proteínas fuera del grupo. A continuación, se evalúa la significación estadística de la asociación entre posiciones que muestran una conservación dependiente de grupo en función de su ajuste a un modelo evolutivo representado por el árbol filogenético asociado al MSA.
Métodos basados en el análisis multivariante del MSA (Sequence Space)	
Sequence Space	Tratamiento vectorial del MSA con Análisis de Componentes Principales (ver texto principal). Requiere la supervisión manual del usuario para definir subfamilias y SDPs
SS-Method	Partiendo de Sequence Space realiza un agrupamiento automático en el espacio de residuos para predecir SDPs. La detección en subfamilias sigue siendo manual.
Métodos que no encajan en las categorías anteriores	
Algoritmo SOM modificado	Utiliza una variante de los mapas autoorganizativos de Kohonen para agrupar las secuencias del alineamiento en subfamilias y detectar sus SDPs a través de los vectores característicos del mapa (*slot vectors*).
k-pax	Utiliza un planteamiento bayesiano para definir simultáneamente subfamilias y SDPs. Comienza definiendo un modelo probabilístico de los patrones presentes en el MSA basado en tres componentes: una partición en un número determinado de subfamilias, el conjunto de posiciones informativas para obtener esa partición y el subconjunto de posiciones informativas que son características de cada una de las subfamilias. A continuación, a partir de la asunción de ciertas probabilidades a priori y mediante un algoritmo de búsqueda estocástica voraz (*greedy stochastic search algorithm*) se persigue resolver las tres componentes del modelo que maximizan su probabilidad a posteriori dado el MSA de partida
SPEL	Evalúa la verosimilitud (*log-likelihood ratio*) de observar en cada posición del MSA su distribución concreta de residuos condicionada al árbol filogenético asociado y a un modelo determinado de sustitución de aminoácidos. A continuación se asigna un p-valor a la verosimilitud de cada posición por contraste estadístico con un modelo aleatorio que, ajustándose al árbol, asume la ausencia de restricciones evolutivas específicas de posición. Las posiciones con un p-valor significativo se consideran las determinantes de la especificidad funcional.
CEO	Algoritmo para establecer una partición en subfamlias en el MSA de modo que se optimice el compromiso entre el número de proteínas en cada subfamilia y el número de SDPs necesario para distinguir unas subfamilias de otras. Para ello se rastrean las particiones en subfamilias arrojadas por un método guiado de agrupación jerárquica determinista. La partición en subfamilias seleccionada es aquella que minimiza una función de contraste entre la conservación diferencial entre las subfamilias de la partición (medida en términos de entropía) y la correspondiente a una partición aleatoria.

Junto a las metodologías anteriores, existen otros desarrollos que parten de una clasificación preestablecida en subfamilias, y por ello se catalogan como "supervisados". Al disponer a priori de las agrupaciones de proteínas, estos abordajes son mucho más sencillos que los anteriores "no supervisados" (**Tabla 2 y 3**). y en esencia tratan de identificar posiciones del MSA con una distribución de los tipos de aminoácidos conservada diferencialmente entre los grupos (SDPs correspondientes a la clasificación impuesta). La mayoría de ellos utilizan para ello medidas basadas en la teoría de la información como la entropía relativa o la información mutua (Hannenhalli y Russell, 2000; Mirny y Gelfand 2002; Kalinina *et al.* 2004; Chakrabarti *et al.* 2007), aunque existen abordajes más simples basados en sistemas "de cuentas" (Capra y Singh 2008) o diseñados para trabajar únicamente con una partición en dos grupos (Pirovano *et al.* 2006; Capra y Singh 2008). Algunos de estos métodos consideran explícitamente la conservación diferencial de las propiedades fisicoquímicas de los aminoácidos (Chakrabarti *et al*. 2007) o incorporan información estructural para detectar residuos agrupados en superficie (Ye et al. 2008). Si bien todos ellos están diseñados para detectar SDPs sobre la agrupación de partida, algunos pueden además identificar posiciones conservadas en un solo grupo aunque no exista conservación de otros tipos de aminoácido en el resto (Chakrabarti et al 2007).

I.11. Importancia de las subfamilias y SDPs en la especificidad funcional en proteínas

Como se comentó anteriormente, la hipótesis de trabajo generalmente aceptada es que la divergencia en secuencia entre las subfamilias de proteínas está gobernada por una divergencia funcional de la que sus posiciones características (SDPs) serían responsables. Numerosos estudios computacionales han mostrado la relación de subfamilias y SDPs con importantes aspectos de la especificidad funcional entre proteínas homólogas.

En el caso de los enzimas y las proteínas que unen pequeños ligandos, esta relación está bien establecida: trabajos a gran escala como los de Madabushi *et al*. (2002), Yao *et al*. (2003), Lichtarge *et al*. (2003), del Sol *et al*. (2003) y Pei *et al*. (2006) junto a otros a mediana escala como los de Landgraf *et al*. (2001), Reva *et al*. (2007) y Kalinina *et al.* (2009), han mostrado de forma robusta la asociación de SDPs con regiones estructurales correspondientes a sitios catalíticos y de unión a pequeños ligando (**Tabla 3**). En el caso particular de los enzimas, la correspondencia explícita de subfamilias con diferentes especificidades catalíticas ha sido también puesta de manifiesto a gran escala (Wicker *et al.* 2001; Brown *et al*. 2007 ; Lee *et al*. 2009), si bien de la mano de métodos especializados en la detección de subfamilias que no detectan SDPs.

Referencia	Método utilizado	Número total de familias estudiadas	Región estructural investigada y número de familias estudiadas		
			Sitios catalíticos y de unión a pequeños ligandos	Interfaces Proteína – Proteína	Sitios de unión a ADN / ARN
Landgraf *et al.* 2001	*3D Cluster Analysis*	35	15	25	6
Madabushi *et al* 2002	ET	38	38	-	-
Yao *et al* 2003	ET	57	57	-	-
Lichtarge *et al* 2003	ET	84	84	-	-
del Sol *et al.* 2003	S-Method MB y SS-Mehtod	303	303	-	-
Pei *et al.* 2006 (*)	SPEL	57	57	-	-
Reva *et al* 2007	CEO	20	10	14	5
Kalinina *et al.* 2009	SDPsite	26	26		

(*) En Pei *et al*. 2006 se evalúa el mismo conjunto de familias que en Yao *et al*. 2003

Tabla I.3. Relación de estudios realizados a mediana/gran escala utilizando diferentes metodologías para la detección de SDPs. Se detalla el número de familias que evaluadas en cada trabajo en función de diferentes tipos de regiones funcionales.

También en casos concretos se ha encontrado una implicación de subfamilias y SDPs en la unión específica a ADN/ARN (Lichtarge *et al*. 1997; Mirny y Gelfand. 2002; Donald y Shakhnovich 2005b; Landgraf *et al.* 2001; Mihalek *et al.* 2004; Reva *et al.* 2007).

De forma importante para esta tesis, la relación entre subfamilias y SDPs con interacciones específicas proteína-proteína se ha indagado para un número considerable de familias concretas (Lichtarge *et al.* 1996a; Innis *et al*. 2000; Mihalek *et al.* 2004; Pupko *et al*. 2002; Bicket *el al*. 2002; Mirny y Gelfand 2002; Bezerin *et al*. 2004; La *et al*. 2005) así como en estudios a pequeña escala (Landgraf *et al.* 2001; Reva *et al.* 2007). Estos estudios han mostrado la participación ocasional de SDPs en interfaces de interacción, sugiriendo la importancia funcional de estas posiciones en la determinación de las interacciones específicas con diferentes proteínas.

La relevancia funcional de subfamilias y SPDs atañe no sólo a las interacciones entre proteínas distintas sino también a la forma de interacción de

los homo-oligómeros. Recientemente, Dayhoff *et al.* (2010) han estudiado la evolución de nueve familias de proteínas en las cuales el modo de homo-oligomerización varía entre sus miembros. Mediante el mapeo de sus diferentes formas de unión en los árboles filogenéticos de cada familia, estos autores han observado que los modos de simetría tienden a estar conservados dentro de las diferentes subfamilias y de forma diferencial entre ellas.

Cabe destacar que la importancia de los SDPs en la determinación de la especificidad funcional que los estudios computacionales han señalado, se ha seguido de validación experimental en una serie importante de casos, relacionados tanto con la especificidad de enzimas (p.ej. Morillas *et al.* 2002) como con interacciones específicas entre proteínas (p.ej. Onrust *et al.* 1997; Bauer *et al.* 1999; Sowa *et al.* 2001; Hernández-Falcón *et al.* 2004 y Juan *et al*. 2005). Así por ejemplo, en Morillas et al. (2002) se identificaron y validaron experimentalmente las posiciones responsables de la inhibición catalítica de las enzimas carnitín-palmitoíl-transferasa y carnitín-octanoíl-transferasa. En Bauer *et al*. (1999) se identificaron los residuos responsables de las diferentes especificidades de interacción con distintos efectores de las proteínas reguladoras Ras y Ral. Las predicciones fueron validadas experimentalmente mostrando que el reemplazo de dos posiciones específicas provoca el intercambio de sus especificidades de unión. Por su parte, Hernández-Falcón *et al.* (2004) y Juan *et al.* (2005) identificación dos residuos críticos en la dimerización del receptor de quimioquina CCR5 que fueron también comprobados experimentalmente.

I.12. Propuesta de un estudio integrado a gran escala de importantes aspectos de la especificidad funcional relacionados con la organización en subfamilias y sus SDPs característicos

Los trabajos que se comentan en el apartado anterior sirven de base para proponer como objetivo principal de esta tesis un estudio integrado en el que se analice a gran escala la relación de subfamilias con patrones de interacción diferencial entre proteínas homólogas así como la implicación de SDPs en interfaces proteína-proteína. Para una mejor comprensión de las señales en secuencia relacionadas con la evolución de la especificidad funcional, conviene que este estudio se haga de forma conjunta con otros importantes aspectos funcionales como los relacionados con la actividad enzimática y la unión a pequeños ligandos, los cuales se sabe que pueden estar íntimamente relacionados con la interacción específica entre proteínas.

Para este estudio es necesario un abordaje metodológico en el que, a partir de un MSA, se definan subfamilias además de SDPs (**Tabla 1**). De entre ellos, el de Sequence Space resulta especialmente idóneo por su capacidad de

explotar la mutua dependencia entre ambas entidades: en un abordaje multivariante las posiciones del MSA determinan la separación de las proteínas y al mismo tiempo esta separación pondera la contribución de las posiciones a esa segregación. Sin embargo la necesidad de supervisión manual en las dos implementaciones disponibles de *Sequence Space* (Casari *et al.* 1995 y del Sol *et al.* 2003) impiden su uso para un estudio a gran escala como el que aquí se propone. Así, como primer objetivo de esta tesis se propone el desarrollo de una metodología de análisis multivariante en el que, a partir de un MSA, subfamilias y SDPs puedan ser detectadas de forma simultánea y coherente.

Por último, este estudio se complementa con el abordaje de situaciones específicas en las que los parecidos relativos en secuencia de las proteínas no se corresponden con sus características funcionales observadas (p.ej. alineamientos estructurales de homólogos remotos). Para predecir residuos funcionales en estas situaciones, se presentan dos metodologías que hacen uso de información funcional conocida, tanto de tipo cuantitativo como en forma de clasificación supervisada.

II. Objetivos

OBJETIVOS

El **objetivo principal** de esta tesis es ahondar en la comprensión de los aspectos funcionales que han modulado la divergencia en secuencia de las familias de proteínas homólogas. Para ello se propone estudiar la asociación de su organización en subfamilias y sus residuos característicos con elementos fundamentales de la especificidad funcional.

Este objetivo principal se compone de los siguientes **objetivos específicos**:

1) El desarrollo de un protocolo automatizado de análisis de familias de proteínas basado en secuencia capaz de detectar su estructura interna en subfamilias y sus residuos diferencialmente conservados de forma coherente y simultánea.

2) El estudio integrado sobre un amplio conjunto de familias de proteínas de la asociación de subfamilias y sus residuos característicos con aspectos funcionales específicos relacionados con la actividad enzimática, la unión diferencial a pequeños ligandos y las interacciones proteína-proteína.

3) El desarrollo de métodos computacionales basados en secuencia capaces de explotar información funcional cuantitativa y clasificaciones supervisadas de familias proteínas para la predicción de los residuos responsables de sus características específicas.

III. Resultados

Parte I
Desarrollo de un método no supervisado basado en análisis multivariante para la detección simultánea de subfamilias de proteínas y sus correspondientes posiciones determinantes de especificidad funcional

En esta parte se presenta el desarrollo de un nuevo método no supervisado para la detección totalmente automatizada de subfamilias de proteínas y sus correspondientes posiciones determinantes de especificidad funcional (SDPs): el método S3det. Este método utiliza como información de partida alineamientos múltiples de secuencias (MSAs). Se califica como no supervisado porque no requiere información funcional externa sino que explota la segregación implícita en el alineamiento. El método S3det es conceptualmente similar al método Sequence Space (Casari *et al.* 1995) presentado en la Introducción, si bien hace uso de herramientas matemáticas esencialmente diferentes y es totalmente automático.

Se describe a continuación el procedimiento implementado en S3det para la determinación automática de subfamilias y SDPs a partir de un MSA. A fin de que esta descripción sea más asequible para un lector no experto, se comienza con una explicación intuitiva remitiendo a las correspondientes subsecciones donde los detalles de cálculo son expuestos formal y exhaustivamente.

III.I.1. El método S3det

El marco metodológico escogido para el análisis de MSAs ha sido el Análisis de Correspondencias Múltiple (MCA) (Greenacre 1984), una técnica estadística de análisis descriptivo multivariante basada en el Análisis de Correspondencias (Greenacre 1984). El MCA puede considerarse como el equivalente del Análisis de Componentes Principales (PCA) cuando se trabaja con datos binarios o cualitativos de manera tal que pueda derivarse para su tratamiento una métrica continua de forma objetiva y coherente (Greenacre, 1984; Lebart *et al.* 1984). Aplicado a un alineamiento múltiple de secuencias, el MCA proporciona una representación vectorial simultánea de proteínas y residuos en un espacio definido por las fuentes de variación independientes en el seno del MSA. Este abordaje permite así identificar grupos de proteínas particularmente vinculadas a determinados grupos de residuos, los cuales serán considerados ulteriormente como subfamilias de proteínas y SDPs respectivamente. En la **Fig. 2** se muestra una representación esquemática del funcionamiento del método S3det que se desarrolla a continuación.

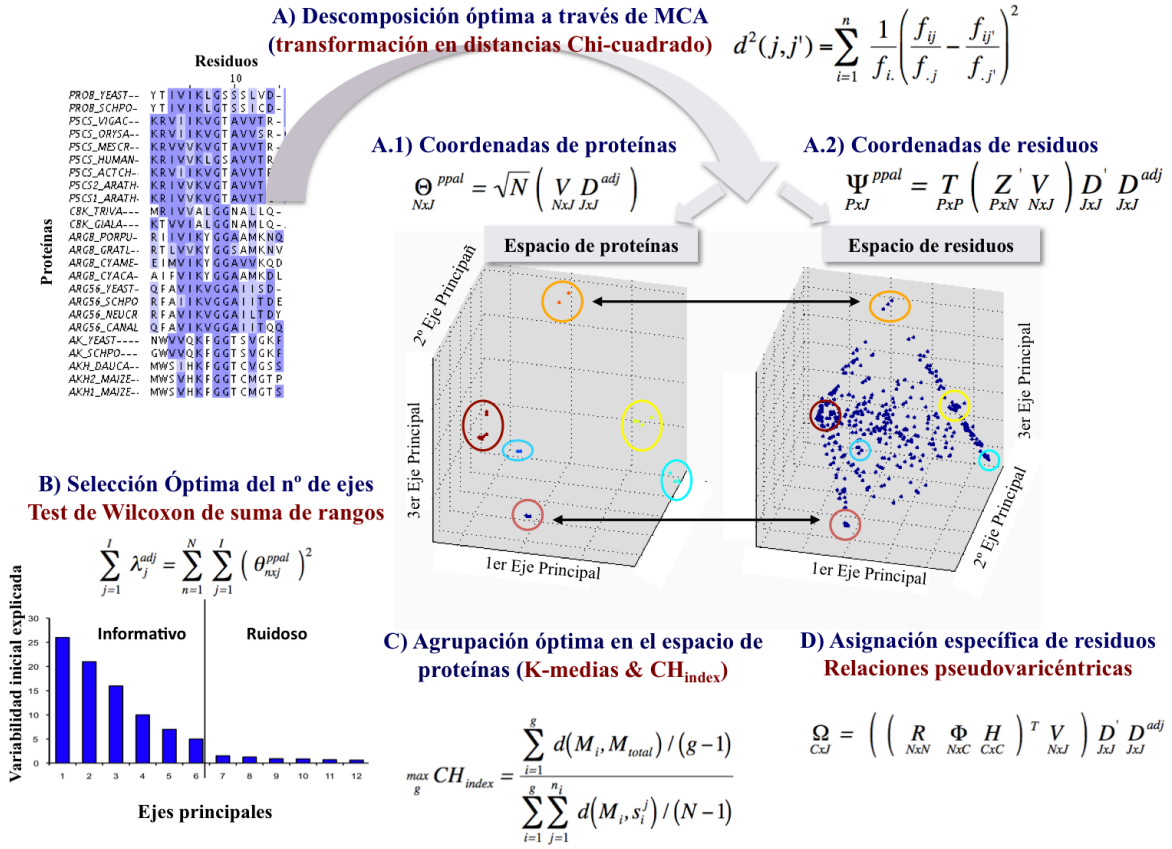


Fig. 2. Representación esquemática del funcionamiento del método S3det (ver texto).

En primer lugar, el MSA se codifica como una matriz binaria (**Fig. 3**; **Sección III.I.1.A**) que implica ya una primera representación vectorial tanto de proteínas como de residuos. El MCA realiza sobre esta matriz una transformación del sistema de coordenadas obteniendo los –así llamados– "ejes principales" (**Fig. 1a**; **sección III.I.1.B**). La principal característica de los "ejes principales" es que están incorrelacionados (son ortogonales). Esta característica es crítica para que las distancias que vayan a calcularse entre los elementos del espacio (proteínas o residuos) tengan sentido† (véase nota al pie).

† Considérese p.ej. un determinado vector $\vec{u}$ de coordenadas (0,0,0) y otro vector $\vec{v}$ de coordenadas (1,1,√2). La distancia euclídea entre $\vec{u}$ y $\vec{v}$ viene dada por d=√ [(1-0)^2 +(1-0) ^{2} + (√2-0) ^{2}]=2. Nótese que se ha asumido que las coordenadas de ambos vectores representan sus coordenadas en un espacio euclídeo tridimensional $\Re^3$ en el que sus 3 ejes ($\vec{i}$, $\vec{j}$ y $\vec{k}$) son ortogonales. No obstante, si en realidad $\vec{u}$ y $\vec{v}$ se encontraran en un espacio euclídeo bidimensional $\Re^2$ en el que sólo los ejes $\vec{i}$ y $\vec{j}$ fueran ortogonales y el tercer eje $\vec{k}$ fuera una combinación lineal cualesquiera de $\vec{i}$ y $\vec{j}$, entonces la distancia entre $\vec{u}$ y $\vec{v}$ vendría dada por d= √ [(1-0)^2 +(1-0) ^{2}]= √2, con independencia de sus coordenadas en el tercer "eje". Se observa

Cada uno de estos ejes explica además una parte de la variabilidad total del MSA y –siendo ortogonales- pueden interpretarse como sus fuentes independientes de variación. A continuación se calculan las coordenadas de proteínas y residuos en los ejes principales (**Fig. 2a.1** y **Fig. 2a.2**; **secciones III.I.1.C** y **III.I.1.D** respectivamente). En esta nueva representación vectorial, las distancias entre cualquier par de proteínas, así como las distancias entre cualquier par de residuos, equivalen a sus distancias Chi-cuadrado. La distancia Chi-cuadrado es especialmente apropiada para tratar con datos binarios/cualitativos, corrigiendo el "peso" que tienen en el análisis en función de la frecuencia con la que aparecen (Peña 2002). Esta es la característica distintiva de MCA respecto a PCA, el cual hace uso, en cambio, de la distancia euclídea. La principal ventaja de utilizar distancias Chi-cuadrado es el hecho de que, en el sistema de coordenadas obtenido, proteínas y residuos pueden compararse directamente en términos de distancias. Esto es así gracias a las relaciones pseudovaricéntricas entre ambos conjuntos de elementos. De acuerdo a estas relaciones, el centro de masas de cualquier grupo de proteínas señala aquellos residuos particularmente asociados a ellas (**sección III.I.1.E**).

El marco provisto por el MCA permite establecer con criterios objetivos un balance óptimo entre la cantidad de variabilidad inicial que será considerada informativa y aquella que será considerada "ruido". Este balance se realiza a través de la cantidad de información que aporta cada uno de los ejes principales. Para determinar qué ejes se consideran relevantes, la información incremental que añaden se evalúa estadísticamente a través del llamado "Test de Wilcoxon de Suma de Rangos" (Miller y Miller 1998) (**Fig. 2b**; **sección III.I.1.F**).

Una vez seleccionados el número de ejes a considerar, se realiza el agrupamiento de proteínas (*clustering*) mediante el algoritmo de k-medias (**Fig. 2c**). El algoritmo se ejecuta sucesivamente para un número predeterminado de grupos, desde dos hasta un cuarto del número de proteínas en el alineamiento (con un máximo de cincuenta). A continuación, se evalúan las diferentes soluciones correspondientes a cada número de grupos prefijado. La solución finalmente escogida será aquella que maximice el llamado CH_{index} (Calinski y Harabasz 1974), el cual mide el ratio de las distancias entre grupos sobre las distancias dentro de grupo (**Fig. 2c**). Se establecen así de forma automática las agrupaciones de proteínas que serán consideradas como las diferentes subfamilias que componen el MSA (**sección III.I.1.G**).

Finalmente, los centros de masas de las agrupaciones de proteínas se emplean para determinar aquellos residuos particularmente asociados a ellas (**Fig 2d**; sección III.I.1.H). Las posiciones determinantes de especificidad (SDPs) se identifican como aquellas posiciones cuyos residuos están especialmente

aquí la necesidad de contar con un espacio vectorial ortogonal (de ejes no correlacionados) en el que el cálculo de distancias entre secuencias y entre residuos sea informativo y no un artefacto.

asociados a los grupos que componen la partición en subfamilias obtenida (**sección III.I.1.I**).

En las siguientes subsecciones se detallan formalmente los cálculos implementados en el método S3det.

III.I.1.A. Codificación binaria de un alineamiento múltiple de secuencias

Dado un alineamiento múltiple de secuencias (MSA) con *N* secuencias y *L* posiciones, se puede derivar una matriz binaria $\underset{NxQ}{W}$ de dimensiones *NxQ* donde *Q=21L* de tal modo que cada posición "*l*" en el MSA inicial se codifica como una categoría disjunta completa con 21 modalidades, correspondientes a los 20 tipos de aminoácidos más la opción de hueco (*gap* en inglés). Así, para una secuencia dada, cada modalidad se representa con un "1" si está presente en ella o con un "0" en caso contrario (**Fig. 3**).

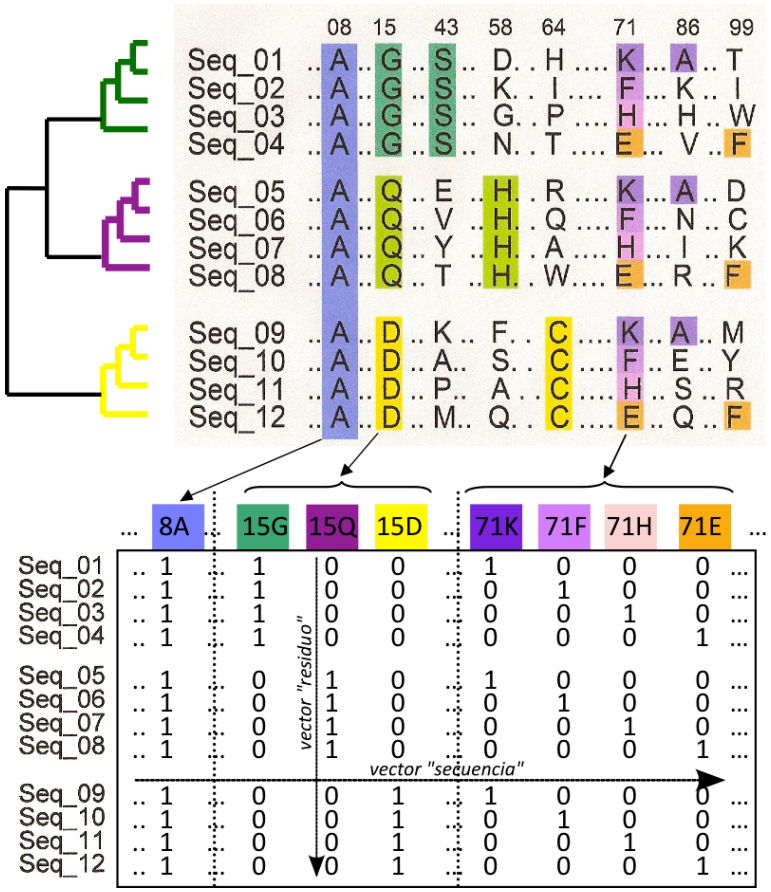


Fig. 3. Codificación binaria de un alineamiento múltiple de secuencias de proteínas. (Arriba) Se representa un alineamiento hipotético de doce secuencias de proteínas. Las relaciones filogenéticas implícitas entre las secuencias se muestran a su izquierda en forma de árbol. **(Abajo)** Codificación del MSA en forma de matriz binaria. En esta matriz, cada columna original del MSA se desglosa en tantas columnas como tipos de aminoácidos diferentes

presenta. Para cada secuencia se asigna un "1" si presenta el aminoácido que se está codificando en esa posición, o un "0" en caso contrario (**Sección III.I.1.A**).

Las columnas en *W* con todos sus elementos nulos se eliminan para la consistencia ulterior del análisis, sin pérdida de generalidad, resultando en una matrix $\underset{NxP}{X}$ de dimensiones *NxP*, donde *P<Q*.

III.I.1.B Determinación de los ejes principales y su varianza explicada

A partir de la matriz $\underset{NxP}{X}$ con término general x_{np} definida en la sección anterior, se definen las siguientes frecuencias:

$$x_{nS} = \sum_{p=1}^{P} x_{np} \; ; \; x_{Sp} = \sum_{n=1}^{N} x_{np} \; ; \; x_{SS} = \sum_{n=1}^{N}\sum_{p=1}^{P} x_{np} \tag{1}$$

$$f_{nS} = x_{nS} / x_{SS} \; ; \; f_{Sp} = x_{Sp} / x_{SS} \; ; \; f_{np} = x_{np} / x_{SS} \tag{2}$$

Sea $\underset{NxP}{Y}$ la matriz de término general $y_{np} = f_{np} / \left(f_{Sp} \sqrt{f_{nS}} \right)$

Sea $\underset{NxP}{Z}$ la matriz de término general $z_{np} = f_{np} / \sqrt{f_{Sp} f_{nS}}$ y $\underset{PxN}{Z^T}$ su transpuesta.

Sea $\underset{NxP}{V}$ la matriz que representa por columnas los vectores principales de ZZ^T.

El espacio generado por los vectores principales de ZZ^T constituye una descomposición ortogonal de la asociación del conjunto de proteínas versus el conjunto de residuos, entre sus fuentes de variación (Peña 2002).

En aras de la consistencia y el incremento de la bondad del ajuste (Greenacre y Blasius 2006), se realiza *a posteriori* la corrección de los valores propios asociados a cada vector propio: Sea λ_i el *i*-avo valor propio no nulo de la matriz ZZ^T. La solución trivial $\lambda=1$ y aquellos valores propios para los que se tenga que $\lambda_i < \left(1/L^2\right)$ -así como sus vectores principales asociados- son descartados del análisis. De este modo, $\underset{NxP}{V}$ resulta en $\underset{NxJ}{V}$, donde *J* es el número de valores propios remanente. El ajuste que se realiza es el siguiente:

$$\lambda_j^{adj} = \left(\frac{L}{L-1} \right)^2 \left(\lambda_j - \frac{1}{L} \right)^2 \tag{3}$$

Así, el porcentaje de la varianza total explicada por cada eje "*i*" se refleja en los valores propios asociados como sigue:

$$\frac{\lambda_i^{adj}}{\sum_{j=1}^{J} \lambda_j^{adj}} \cdot 100 \qquad (4)$$

III.I.1.C. Obtención de las coordenadas de las proteínas del MSA sobre los ejes principales

Sea $\underset{NxN}{R}$ la matriz diagonal de dimensiones *NxN* de término general $r_{nn} = \frac{1}{\sqrt{f_{nS}}}$

Sea $\underset{PxP}{T}$ la matriz diagonal de dimensiones *PxP* de término general $t_{pp} = \frac{1}{\sqrt{f_{Sp}}}$

Sea $\underset{JxJ}{D}$ la matriz diagonal de dimensiones *JxJ* de término general $d_{jj} = \sqrt{\lambda_j}$ y $\underset{JxJ}{D'}$ su inversa

Sea $\underset{JXJ}{D^{adj}}$ la matriz diagonal de dimensiones *JxJ* de término general $d_{jj}^{adj} = \sqrt{\lambda_j^{adj}}$

La proyección de las *N* secuencias en el espacio generado por los $\underset{NxP}{V}$ vectores principales de ZZ^T viene dada por

$$\underset{NxJ}{\Theta^{ppal}} = \sqrt{N} \left(\underset{NxJ}{V} \underset{JxJ}{D^{adj}} \right) \qquad (5)$$

donde el término general θ_{nj}^{ppal} de $\underset{NxJ}{\Theta^{ppal}}$ equivale a la "coordenada principal" de la proteína "*n*" en el eje principal "*j*". Así mismo, las llamadas "coordenadas estándar" se calculan como sigue:

$$\underset{NxJ}{\Theta^{stdr}} = \sqrt{N} \left(\underset{NxJ}{V} \underset{JxJ}{D'} \underset{JxJ}{D^{adj}} \right) \qquad (6)$$

III.I.1.D. Obtención de las coordenadas de los residuos del MSA sobre los ejes principales

De forma análoga, la proyección de los *P* residuos en el espacio generado por los $\underset{NxP}{V}$ vectores principales de ZZ^T viene dada por

$$\underset{PxJ}{\Psi^{ppal}} = \underset{PxP}{T} \left(\underset{PxN}{Z'} \underset{NxJ}{V} \right) \underset{JxJ}{D'} \underset{JxJ}{D^{adj}} \qquad (7)$$

cuyo término general ψ_{pj}^{ppal} de $\underset{PxJ}{\Psi}^{ppal}$ equivale a la "coordenada principal" del residuo "*p*" en el eje principal "*j*".

III.I.1.E. Relaciones pseudovaricéntricas entre las coordenadas de las proteínas y las coordenadas de los residuos

Sea *C* un numero cualesquiera de características adicionales para las cuales las *N* proteínas del MSA pueden ser evaluadas en términos binarios. Esto es, se puede construir una matriz binaria $\underset{NxC}{\Phi}$ de término general φ_{nc} donde $\varphi_{nc} = 1$ si la proteína "*n*" presenta la característica "*c*", ó $\varphi_{nc} = 0$ en caso contrario. En términos lógicos:

$$\varphi_{nc} \begin{cases} \varphi_{nc} = 1 \quad \Leftrightarrow \quad n \in c \;\; ; \;\; n\{1,..N\} \;\; , \;\; c\{1,..C\} \;\; ; \;\; \forall c \; \exists n \;\; / \;\; \varphi_{nc} = 1 \\ \varphi_{nc} = 0 \quad \Leftrightarrow \quad n \notin c \end{cases} \tag{8}$$

Sea $\underset{CxC}{H}$ la matriz diagonal de dimensiones *CxC* de término general $h_{cc} = 1/\sum_{n=1}^{N} \varphi_{nc}$

Debido a las llamadas "relaciones pseudovaricéntricas" que caracterizan el MCA, se demuestra que las "coordenadas principales" resultantes de proyectar una característica adicional "*c*", representada en forma de columna en $\underset{NxC}{\Phi}$, como un elemento suplementario en el espacio generado por los *J* vectores principales $\underset{NxP}{V}$ de ZZ^T junto con los *P* residuos, equivalen a las coordenadas del centro de masas de las "n" proteínas que presentan la característica "*c*" calculadas a partir de sus coordenadas estándar ϕ_{nj}^{stdr}. Las coordenadas principales ω_{cj}^{ppal} de la característica "*c*" sobre los *J* ejes principales vienen dadas por la siguiente expresión:

$$\underset{CxJ}{\Omega} = \left(\left(\underset{NxN}{R} \; \underset{NxC}{\Phi} \; \underset{CxC}{H} \right)^T \underset{NxJ}{V} \right) \underset{JxJ}{D}^{'} \; \underset{JxJ}{D}^{adj} \tag{9}$$

Podemos resumir los procedimientos seguidos hasta este punto como sigue: el MCA realiza una descomposición ortogonal de las fuentes de variación implícitas en el MSA inicial. Estas fuentes se representan a través de cada uno de los ejes principales (o vectores principales) codificados en $\underset{NxP}{V}$, las cuales pueden ser priorizadas a través de sus valores propios asociados ordenándolos de forma decreciente. Las proteínas y posiciones del MSA se pueden

representar entonces en este espacio a través de sus coordenadas principales: $\underset{NxJ}{\Theta}^{ppal}$ y $\underset{PxJ}{\Psi}^{ppal}$ respectivamente. Como una de las características fundamentales del MCA, el valor propio asociado a cada eje equivale a la varianza de las coordenadas principales en ese eje, tanto en el caso de las proteínas como en el de los residuos. En el espacio generado, la distancia euclídea calculada en base a coordenadas principales -entre cualquier par de proteínas o entre cualquier par de residuos- equivale a una distancia Chi-cuadrado. Las distancias Chi-cuadrado son especialmente apropiadas para comparar individuos descritos a través de variables cualitativas (Peña, 2002). El espacio generado por el MCA permite también la comparación de proteínas contra residuos (y a la inversa) a través de las llamadas "relaciones pseudovaricéntricas", esto es: el centro de masas basado en "coordenadas estándar" de un conjunto dado de proteínas equivale a las "coordenadas principales" de un residuo (o cualquier otra característica binaria) que se ajuste de forma exacta al patrón de presencia/ausencia en esas proteínas determinadas. De este modo, un conjunto dado de proteínas puede representarse, como tal entidad, en el espacio de coordenadas principales de residuos y ser así directamente comparado con estos en términos de distancias euclídeas (que equivalen a su vez a distancias chi-cuadrado sobre la matriz de asociación original).

III.I.1.F. Selección del número de ejes principales informativos

Una vez descompuesta la variabilidad total presente en el MSA inicial en fuentes de variación ortogonales y ordenadas estas en orden decreciente de importancia, el siguiente paso consiste en evaluar cuántos de estos ejes deben considerarse relevantes. En este trabajo se aborda esta selección de forma analítica y objetiva mediante el empleo de un test estadístico no paramétrico: el llamado "Test de Wilcoxon de Suma de Rangos" (Miller y Miller 1998). El objetivo es calcular la significación estadística del aumento de información que se incorpora al análisis cuando se considera una nueva dimensión. El número *I* de ejes considerados relevantes se calcula como sigue:

(10)

$$I = \begin{cases} max\ i \in [2,J]\ /\ Prob_{Wilcoxon}\left(\left[\sum_{j=1}^{i-1} \left(\theta_{nxj}^{ppal} \right)^2, \forall n \in N \right] < \left[\sum_{j=1}^{i} \left(\theta_{nxj}^{ppal} \right)^2, \forall n \in N \right] \right) < 0.01 \\ 1 \end{cases}$$

Por razones prácticas en el tiempo de cálculo, en este trabajo se ha obligado a *I* a ser menor o igual que "10". La comparación de las distribuciones indicada en la expresión anterior se fundamenta en la relación intrínseca entre la variabilidad explicada por un espacio definido por un número *I* de dimensiones y

la suma de los valores propios asociados que lo constituyen, donde

$$\sum_{j=1}^{I} \lambda_j^{adj} = \sum_{n=1}^{N} \sum_{j=1}^{I} \left(\theta_{nxj}^{ppal} \right)^2 \quad (11)$$

El anterior procedimiento de selección del número óptimo de dimensiones a considerar puede interpretarse como una forma analítica de establecer el equilibrio óptimo entre la información de partida que ha de ser considerada relevante y aquella que puede considerarse "ruidosa".

III.I.1.G. Agrupamiento automático de las proteínas representadas en los ejes principales seleccionados para el establecimiento de subfamilias

El agrupamiento de proteínas se realiza en este trabajo de forma automática a través del algoritmo de k-medias implementado en De Hoon *et al*. (2004). Este se aplica sobre las coordenadas principales de la proteína "*n*" en el eje principal "*j* representadas por Θ_{NxJ}^{ppal} y se ejecuta para un intervalo de número de grupos "*g*" preestablecido. Entre las soluciones aceptadas, se considera óptima aquella que maximice el índice de agrupamiento CH_{index} (Calinski y Harabasz 1974). Este índice mide el ratio de la media de las desviaciones simples inter-grupos sobre la media de las desviaciones simples intra-grupo. Así, la solución óptima del agrupamiento de secuencias será aquella ofrecida por el algoritmo de k-medias anterior para un número *G* seleccionado como sigue:

$$G = g \in \left[2\,,\ min\left(\frac{N}{4}; 50 \right) \right] \quad / \quad \max_g CH_{index} = \frac{\sum_{i=1}^{g} d\left(M_i, M_{total}\right) / \left(g-1\right)}{\sum_{i=1}^{g} \sum_{j=1}^{n_i} d\left(M_i, s_i^j\right) / \left(N-1\right)} \quad (12)$$

donde: *g* es el número de grupos; *N* es el número total de secuencias; n_j es el número de secuencias en el grupo *i* ; M_i es el centro de masas del grupo *i* ; M_{total} es el centro de masas del conjunto total de secuencias (el cual, en el marco del MCA, equivale al origen de coordenadas "*0*"), s_i^j representa a la secuencia *j* del grupo *i*, y *d(_,_)* representa la distancia euclídea calculada sobre la base de las coordenadas codificadas en Θ_{NxJ}^{ppal} . Cuando *N/4* no es una división exacta, se redondea al entero superior. La división de las N proteínas en el número óptimo de grupos *G* será considerada como la composición de subfamilias de proteínas del MSA.

III.I.1.H. Asignación de residuos a subfamilias de proteínas

Cada subfamilia del MSA puede considerarse una característica adicional para la cual las *N* proteínas se pueden evaluar en términos binarios, esto es: con un "1" si la proteína "n" pertenece a esa subfamilia, o con un "0" en caso contrario. De este modo se llega de forma natural a la matriz general de características adicionales $\underset{NxC}{\Phi}$ definida anteriormente. Construimos aquí las *C* características adicionales a partir de todas las posibles formas en que un número γ de subfamilias puede considerarse de forma conjunta de entre el número óptimo *G* previamente establecido y sin consideración de orden, esto es: $\binom{G}{\gamma}$ desde *γ=1* a *γ=G*. Así, el número total de características adicionales C a considerar vendrá dado por la expresión:

$$C = \sum_{\gamma=1}^{\gamma=G} \binom{G}{\gamma} = \sum_{\gamma=1}^{\gamma=G} \frac{G!}{(G-\gamma)!\,\gamma!} \tag{13}$$

Tal como se desarrolló anteriormente, una vez definida la matriz $\underset{NxC}{\Phi}$ las coordenadas principales de cada característica "*c*" en cada uno de los ejes principales "*j*" viene dada por $\underset{CxJ}{\Omega}$, y estas coordenadas se pueden comparar directamente a través de distancias euclídeas contra las coordenadas principales $\underset{PxJ}{\Psi^{ppal}}$ de los residuos. Se puede así compartimentar los *P* residuos asignándolos unívocamente a las *C* agrupaciones de secuencias descritas en $\underset{NxC}{\Phi}$, de modo que cada residuo "*p*" se atribuya a su agrupación se proteínas "*c*" más cercana, esto es:

$$p \in c \Leftrightarrow \min_{c} \; d\left(\overrightarrow{\omega_p}, \overrightarrow{\psi_c^{ppal}} \right) \tag{14}$$

III.I.1.I. Selección de Posiciones Determinantes de Especificidad (SDPs)

Cada conjunto de residuos "*p*", asociado unívocamente a cada agrupación de proteínas "c", se ordena de forma creciente en base a sus distancias respectivas, lo que equivale a ordenar el ajuste de cada residuo en su correspondiente agrupación. Los residuos correspondientes a huecos (*gaps*) se descartan en este momento del análisis de modo que no influyen en los rangos a considerar. Para cada agrupación "*c*", se considera el subconjunto *c'* de posiciones *p'* con los mejores rangos, definido por un límite $\beta \in [0,1]$ que se establece como una fracción del número total de elementos de "c", esto es:

(15)

$$c' \subset c \;/\; \forall c'' \subset c' \;/\; |c''| = \beta \cdot |c| \;:\; max\ rank\left(\forall p'' \in c''\right) = max\ rank\left(\forall p' \in c'\right) < min\ rank\left(\forall p \in \overline{c'}\right)$$

Cuando $\beta \cdot |c|$ no es una división exacta, se redondea al entero superior.

Sea $\underset{LxC}{\Delta}^{ppal}$ una matriz binaria de término general δ_{lc} definido como sigue:

$$\delta_{lc} \begin{cases} \delta_{lc} = rank(p) \;\Leftrightarrow\; p \in c' \subset c \;\wedge\; p \in l \\ \delta_{jc} = 0 \end{cases} \tag{16}$$

El subconjunto de posiciones del MSA que se consideran Posiciones Determinantes de Especificidad (SPDs) se define como aquellas posiciones "*l*" cuyos residuos muestran un rango medio menor o igual que α bajo una partición disjunta completa de las *N* proteínas del alineamiento, esto es:

$$l \in SDPs \Leftrightarrow \begin{cases} \forall k \in C \;/\; \delta_{lk} > 0 \;:\; k \in K \;\wedge\; \sum_k \overrightarrow{\varphi_k} = \vec{1} \;\wedge\; \dfrac{\sum_k \delta_{lk}}{|K|} \le \alpha \\ \forall\, i, j \in C \;/\; \delta_{li} > 0 \;\wedge\; \delta_{lj} > 0 \;\wedge\; i \neq j \;:\; \overrightarrow{\varphi_i} \cdot \overrightarrow{\varphi_j} = 0 \end{cases} \tag{17}$$

En este trabajo se ha establecido α=*10*, β=10 y desestimado cualquier $c \in C \;/\; \left|\overrightarrow{\varphi_c}\right| \le 3$, tanto en la definición de partición disjunta completa como en sus residuos c' asociados.

III.I.2. Ejemplos del funcionamiento de S3det y de su aplicación al estudio de la especificidad funcional en familias de proteínas

III.I.2.A. Familia de aminotransferasas de clase III

El procedimiento seguido para determinar de forma automática y simultánea las subfamilias de proteínas en el seno de un MSA y sus SDPs correspondientes se representa esquemáticamente en la **Fig. 4** a través de los resultados obtenidos para la familia de aminotransferasas de clase III (código Pfam PF00202; **Métodos Sección VI.I**).

El alineamiento obtenido comprende 27 proteínas con las siguientes especificidades enzimáticas (los códigos entre paréntesis corresponden a códigos EC de la *Enzyme Comission*; **Introducción Sección I.2**):

- Acetil-ornitín-aminotransferasa (EC 2.6.1.11), que cataliza la transferencia de un grupo amino desde acetil-ornitina a alfa-ketoglutarato resultando en N-acetilglutámico-5-semialdehído y ácido glutámico.
- Ornitín-aminotransferasa (EC 2.6.1.13), que cataliza la transferencia de un grupo amino desde ornitina a alfa-ketoglutarato resultando en glutámico-5-semialdehído y ácido glutámico.
- 4-aminobutirato-aminotransferasa (EC 2.6.1.19; GABA transaminasa), que cataliza la transferencia de un grupo amino desde GABA a alfa-ketoglutarato resultando en succinato-semialdehído y ácido glutámico.
- Alanín-glioxilato aminotransferasa (EC 2.6.1.44), que cataliza la transferencia de un grupo amino desde L-alanina a glioxilato resultando en piruvato y glicina.
- Glutamato-1-semialdehído aminotransferasa (EC 5.4.3.8) que cataliza la transferencia de un grupo amino intramolecularmente en (S)-4-amino-5-oxopentanoato resultando en 5-aminolevulinato.

El tratamiento mediante MCA del MSA original proporciona una representación vectorial en sendos espacios ortogonales, el de proteínas y el de residuos (representados en la **Fig. 4** utilizando sus tres ejes más informativos). Sobre el espacio de proteínas se realiza un proceso de agrupamiento automático a partir del cual se definen las subfamilias de proteínas. En la figura, los códigos EC a continuación del nombre de las secuencias permiten comparar su correspondencia con las subfamilias encontradas. Como se ve, el agrupamiento automático reproduce las clases enzimáticas, si bien dos de ellas – ECs 2.6.1.11 y 2.6.1.44- se encuentran contenidas en una única subfamilia.

La transformación del MSA inicial a través del MCA, acompañada de la selección del número de ejes principales informativos, implica una transformación de las distancias relativas entre las secuencias (**Sección III.I.1**). Esta diferencia se origina a partir del filtrado de la variabilidad inicial implementado en S3det de tal modo que sólo se retiene la parte considerada informativa en términos de su contribución a la estructura general del alineamiento múltiple (**Sección III.I.1.F**). En la **Fig. 5a** y **5b** se ilustra esta trasformación de distancias relativas mediante la comparación del árbol filogenético[‡] del MSA original y el árbol obtenido a partir de las distancias MCA[§]. Se observa en este último que las subfamilias encontradas presentan una menor variabilidad intra-grupo y una mayor variabilidad inter-grupo que en el primero. Así, por ejemplo, las secuencias del tipo ARGD y AGT (ECs 2.6.1.11 y 2.6.1.44

[‡] Obtenido mediante el método de unión de vecinos (*neighbour-joining*) implementado en el software Phylip (versión 3.69; Felsenstein 1989) usando la matriz de sustitución Blosum62.
[§] Obtenido mediante el método UPGMA *(Unweighted Pair Group Method with Arithmetic Mean*) implementado en el software Phylip (versión 3.69; Felsenstein 1989).

respectivamente), cuyo agrupamiento es dudoso en el árbol inicial, aparecen ahora claramente agrupadas. En general, se puede observar una coherencia global entre la estructura topológica de ambos árboles, siendo más evidente el establecimiento de agrupaciones sobre las distancias MCA que sobre el MSA original.

En S3det, las posiciones del MSA características de la segregación en subfamilias se clasifican simultáneamente junto a estas. La **figura 4** (panel central, abajo) ilustra el espacio de residuos resultante de la descomposición generada por el MCA para la familia de aminotransferasas de clase III. Puede observarse aquí la equivalencia que -en este marco metodológico- poseen ambos espacios, el de proteínas y el de residuos, donde aquellos que se toman como SDPs de las subfamilias se localizan en regiones idénticas a estas.

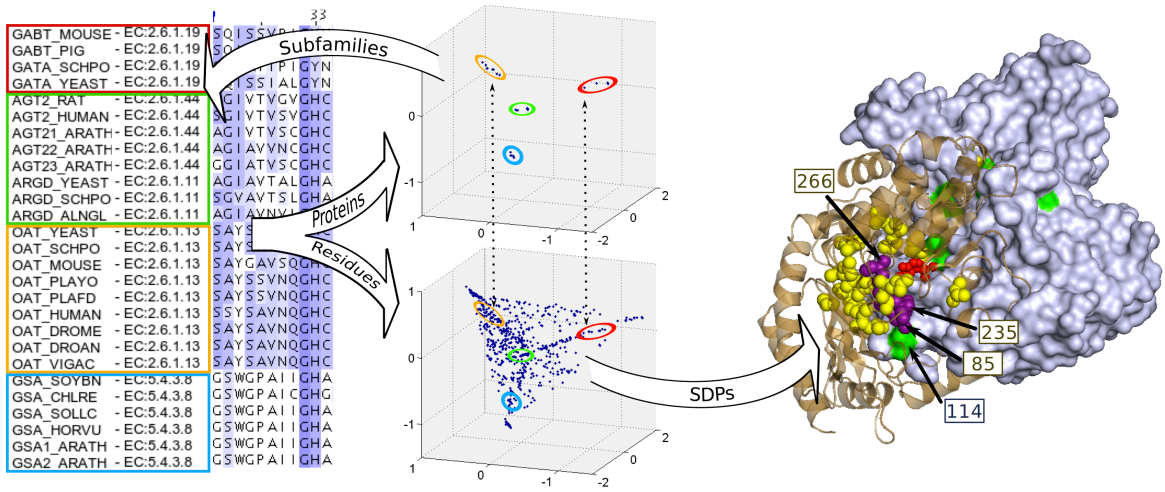
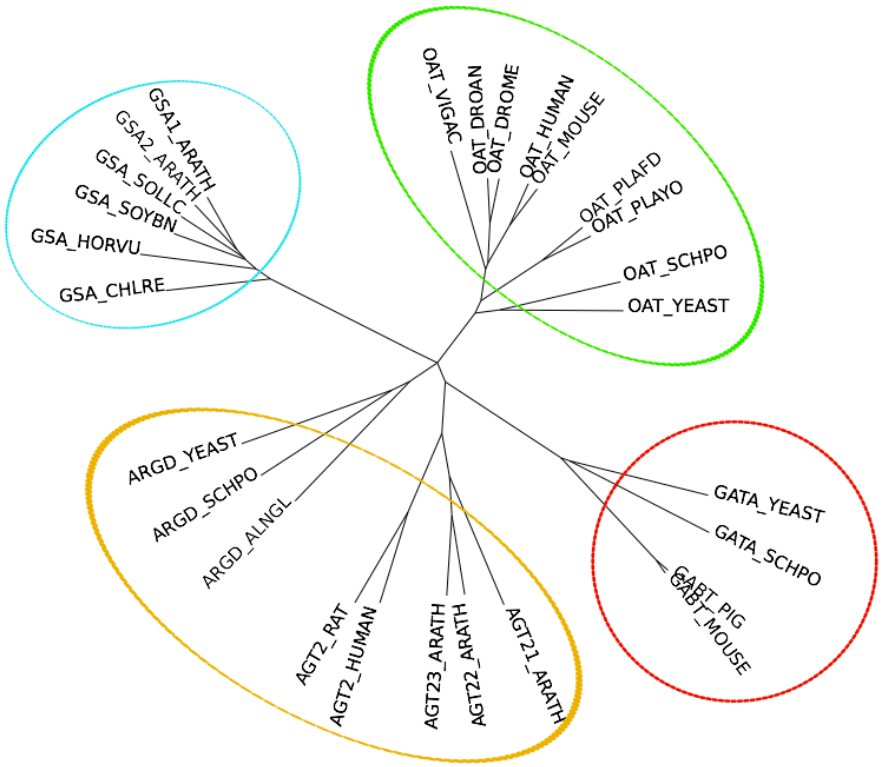


Fig. 4. Representación esquemática de la aplicación del método S3det a la familia de aminotransferasas de clase III (Código Pfam PF00202)**. (Izquierda)** Fragmento del alineamiento múltiple de la familia donde las proteínas aparecen identificadas por su código UniProt ID (Métodos Sección VI.II.3) seguido por el código enzimático EC correspondiente. **(Centro)** Representación vectorial –mediante puntos en el correspondiente espacio- de las proteínas (arriba) y los residuos (abajo) en los espacios obtenidos a través del MCA a partir del MSA inicial. Los espacios se representan en la figura a través de los tres ejes más informativos. En el espacio de proteínas (arriba) se muestran las agrupaciones resultantes del algoritmo de k-medias (Resultados Sección III.I.1.G). El código de colores utilizado es el mismo que el de los recuadros que agrupan los identificadores de proteínas en el MSA (izquierda) representando las subfamilias encontradas. La translación del centro de masas de las agrupaciones anteriores permite establecer los conjuntos de residuos en el espacio respectivo (abajo) a partir de los cuales se determinan las Posiciones Determinantes de Especificidad (SDPs). **(Derecha)** Representación de la estructura homodimérica de la ornitín-aminotransferasa de humanos (*ornithine aminotransferase*; código PDB 1oat) unida a Piridoxal-5'-fosfato (*Pyridoxal-5'-phosphate*; representado en esferas de color rojo). Las dos subunidades del complejo se representan en lazos (*cartoons*) de color marrón y en superficie (*surface*) de color gris. Las SDPs predichas por el método S3det se representan en esferas de color amarillo/violeta y en superficie de color verde. La figura se generó con el programa Pymol (pymol.sourceforge.com).

a)



b)

Fig. 5. Filogramas radiales de la familia de aminotransferasas de clase III obtenidos a partir de diferentes matrices de distancias. (a) Árbol obtenido mediante el método de unión de vecinos (*neighbour-joining*) usando la matriz de sustitución Blosum62 a partir del MSA original. **(b)** Árbol obtenido mediante el método UPGMA aplicado sobre las distancias MCA en el espacio de proteínas. En ambos árboles se resaltan las agrupaciones obtenidas por S3det utilizando el mismo código de colores que en la figura 2.

En la **Fig. 6** se muestran las posiciones predichas como SDPs en el alineamiento de esta familia. Se observa cómo el perfil de aminoácidos de estas posiciones sigue un patrón de conservación diferencial entre subfamilias. Como se explicó en la **Sección III.I.1.H**, este patrón de conservación diferencial puede implicar a las subfamilias de forma independiente (véase p.ej. las posiciones 82 y 89) o en combinaciones entre sí (p.ej. las posiciones 233 o 291). Si bien el tipo de aminoácido está por lo general muy conservado dentro de subfamilia, se observa una mayor variabilidad en el caso del agrupamiento correspondiente a las secuencias ARGD y AGT (ECs 2.6.1.11 y 2.6.1.44). En este grupo se encuentran desde posiciones totalmente conservadas (p.ej. la 291, común a ambos ECs), posiciones semiconservadas (p.ej. la 235, diferencial entre las ARGDs y las AGTs) y posiciones variables (p.ej. la 92). Así, la heterogeneidad funcional de este agrupamiento (ECs 2.6.1.11 y 2.6.1.44) está igualmente reflejada en las SDPs predichas.

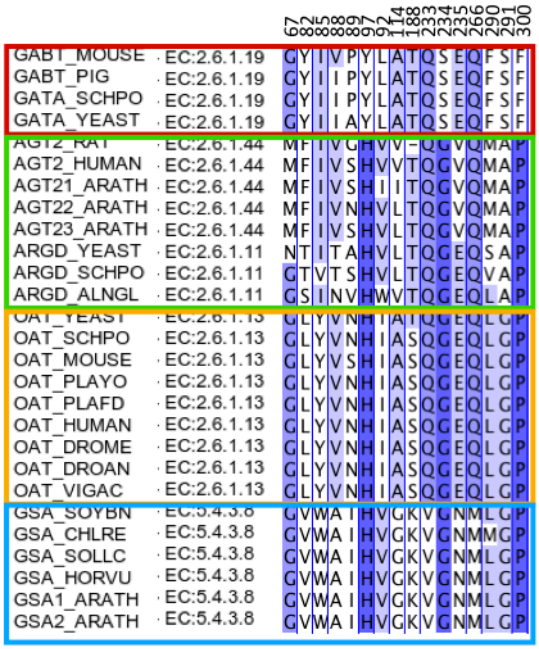


Fig. 6. Posiciones Determinantes de Especificidad (SDPs) predichas por S3det para la familia de aminotransferasas de clase III. La representación de las posiciones no implica aquí su contigüidad en secuencia y están extraídas del alineamiento Pfam (PF00202; Métodos Sección V.I). Las posiciones están enumeradas conforme a los residuos correspondientes a la estructura PDB 1oat. Las subfamilias predichas por S3det se resaltan mediante rectángulos utilizando el mismo código de colores que en las figuras 4 (izquierda) y 5.

En la **Fig. 4** (panel derecho) y la **Fig. 7** se muestra el mapeo de las SDPs predichas sobre la estructura homodimérica de la ornitín-aminotransferasa de humanos (PDB 1oat) tomada como representante de las aminotransferasas de clase III presentes en el alineamiento (**Métodos Sección VI.III.3**). Se observa

que la mayoría de SDPs mapea en la región de interacción entre ambas cadenas y cerca del sitio de unión del ligando.

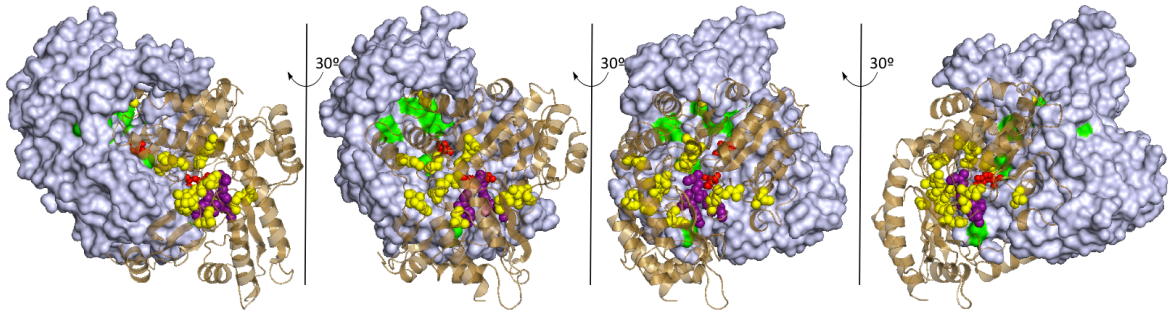


Fig. 7. Diferentes vistas del mapeo de las SDPs predichas para la familia de aminotransferasas de clase III sobre la estructura homodimérica de la ornitín-aminotransferasa de humanos (*ornithine aminotransferase*; código PDB 1oat) unida a Piridoxal-5'-fosfato. La representación sigue los mismos patrones utilizados en la Fig. 4 (derecha).

Entre los SDPs encontrados se incluyen 2 posiciones (85 y 235 en PDB 1oat; **Fig. 4**) cuya implicación en la determinación de la especificidad del sustrato que catalizan se ha demostrado experimental y estructuralmente a través de mutaciones simples (Markova *et al.* 2008). La posición 235 ejerce su influencia sobre la disposición estructural del cofactor Piridoxal-5'-fosfato mediante una red de puentes de hidrógeno que implica a la posición 266 (**Fig. 4**), la cual es también otra SDP predicha. Por otra parte, el residuo 85 de ambas cadenas en el PDB 1oat -a pesar de no estar en contacto directo con el ligando- parece ser uno de los principales condicionantes de esta especificidad. De modo interesante, esta posición se encuentra en contacto con otra SDP (la 114 de la otra cadena; **Fig. 4**) localizada en la interfaz de homodimerización. Esta observación apunta a la posibilidad de que diferentes aminoácidos en posiciones que atañen a la región de interacción pudieran estar modulando la interacción con el ligando y por tanto condicionando la especificidad de sustrato.

III.I.2.B. Familia de factores de transcripción E2F/TPD

Se detalla a continuación los resultados obtenidos para los factores de transcripción de la familia E2F/TDP (código Pfam PF02319; **Métodos Sección VI.I**) cuyos miembros juegan un papel central en la expresión de genes implicados en el ciclo celular (Zheng *et al.* 1999). Este caso ilustra, por una parte, la relación entre subfamilias de proteínas e interactores diferenciales y, por otra, la relación entre sus SDPs y las interfaces correspondientes.

a)

	21	26	32	43	55	61	62	67	E2F4_HUMAN PDB 1cf7 Chain A
	70	75	78	90	120	126	127	132	TDP2_HUMAN PDB 1cf7 Chain B
E2F2_HUMAN	S	T	L	L	K	I	T	G	
E2F3_HUMAN	S	T	L	L	K	I	T	G	
E2F1_CHICK	S	T	L	L	K	I	T	G	
E2F1_HUMAN	S	T	L	L	K	I	T	G	
E2F_DROME	S	T	L	L	K	I	T	G	
E2F4_HUMAN	S	T	L	L	K	I	T	G	
E2F6_HUMAN	S	T	L	L	K	I	T	G	
TDP1_CAEEL	G	S	V	Y	R	A	L	A	
TDP_DROME	G	S	V	Y	R	A	L	A	
TDP2_HUMAN	G	S	V	Y	R	A	L	A	
TDP1_HUMAN	G	S	V	Y	R	A	L	A	
TDP2_MOUSE	G	S	V	Y	R	A	L	A	

b)

	Q14186	Q14188	Q01094	Q14209	O00716	Q16254	Q15329	Q13547	P35232	P38398	P19838	P20226	P28749	P06400
Q14186			+	+		+	+		-	-	-	-	+	+
Q14188			+	+	+	+			-				+	+
Q01094	+	+				-	-	+	+	+	+	+		+
Q14209	+	+				-			-	-		-		+
O00716	+					-	-		-	-		-		+
Q16254	+	+	-	-	-		-	+			-	-	+	+

c)

Swiss AC	Swiss ID	*Name*
Q14186	TFDP1_HUMAN	Transcription factor Dp-1
Q14188	TFDP2_HUMAN	Transcription factor Dp-2
Q01094	E2F1_HUMAN	Transcription factor E2F1
Q14209	E2F2_HUMAN	Transcription factor E2F2
O00716	E2F3_HUMAN	Transcription factor E2F3
Q16254	E2F4_HUMAN	Transcription factor E2F4
Q15329	E2F5_HUMAN	Transcription factor E2F5
Q13547	HDAC1_HUMAN	Histone deacetylase 1
P35232	PHB_HUMAN	Prohibitin
P38398	BRCA1_HUMAN	Breast cancer type 1 susceptibility protein
P19838	NFKB1_HUMAN	Nuclear factor NF-kappa-B p105 subunit
P20226	TBP_HUMAN	TATA-box-binding protein
P28749	RBL1_HUMAN	Retinoblastoma-like protein 1
P06400	RB_HUMAN	Retinoblastoma-associated protein

d)

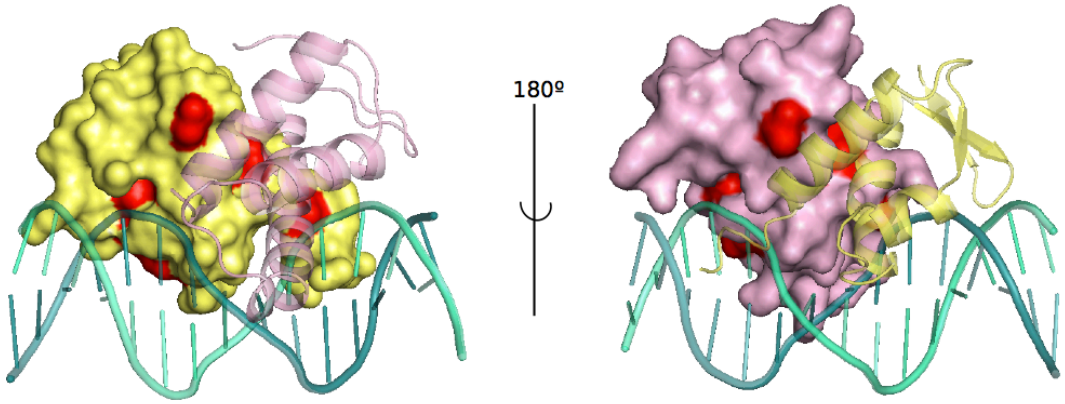


Fig. 8. Resultados obtenidos para la familia de factores de transcripción E2F/TDP (Pfam PF02319)**. (a)** Proteínas contenidas en esta familia (identificadores Uniprot) agrupadas según las subfamilias obtenidas por S3det (cuadro verde para las E2F y amarillo para las TDP) junto a las posiciones del MSA predichas como SDPs. Las posiciones se enumeran conforme a los residuos correspondientes a ambas cadenas de la estructura PDB 1cf7 del heterodímero de humanos E2F4 (cadena A) y TDP2 (cadena B). **(b)** Tabla de interactores positivos y negativos (listados por columnas) de las proteínas de humanos presentes en la familia E2F/TDP (listadas por filas). Las proteínas se identifican por su código de entrada Uniprot. El código de colores utilizado es el siguiente: amarillo (factores de transcripción TDP), verde (los E2F) y salmón (el resto de proteínas interactoras). **(c)** Equivalencia entre identificadores Uniprot (Swiss ID), códigos de entrada en Uniprot (Swiss AC) y nombre de las proteínas listadas en la tabla (b). **(d)** SDPs (representados en superficie con color rojo) mapeados sobre la estructura PDB 1cf7 correspondiente al heterodímero de humanos E2F4 (en amarillo) y TDP2 (en rosa) unida a ADN.

En esta familia se obtuvieron 2 subfamilias que separan de forma clara las proteínas de tipo E2F y las de tipo TDF (**Fig. 8a**). Las proteínas TDP actúan como factores de transcripción dependientes de E2F, mientras que los E2F pueden hacerlo como homodímeros o como heterodímeros (Zheng *et al.* 1999). Los heterodímeros E2F/TDF muestran mayor eficiencia en la unión a ADN (Wu *et al.* 1995), estimulando la transcripción dependiente de E2F. El significado biológico de esta división puede de hecho observarse en la lista completa de interactores (tanto positivos como negativos) obtenida para las proteínas de humanos de esta familia (**Fig. 8b** y **8c**), en la que se refleja la capacidad de los factores de transcripción E2F de formar heterodímeros con las proteínas TDP.

La **Fig. 8d** muestra el mapeo de las SDPs predichas por S3det para esta familia sobre la estructura heterodimérica E2F4/TDP2 humana unida a ADN (PDB 1cf7). Las SDPs se localizan principalmente en la superficie de interacción entre ambas cadenas, complementando a otras SDPs que se encuentran en contacto con el ADN. Se intuye claramente el rol de las SDPs en la determinación del complejo heterodimérico, esto es, en la determinación de una interacción específica entre ambas proteínas.

III.I.3. Implementación de S3det en un software C/C++ distribuíble y alojamiento en el servidor Treedet

El protocolo descrito para la definición automática de subfamilias y SDPs en un MSA se ha implementado en un software escrito en lenguaje C/C++. El programa es de libre uso para fines académicos. El paquete de distribución incluye el código fuente acompañado de archivos de ayuda para su instalación, configuración de variables locales, compilación de ejecutables, opciones de uso y manuales para la interpretación de la salida. Así mismo se incluyen diferentes ejemplos de inputs y sus correspondientes outputs que permiten la

comprobación del correcto funcionamiento del programa una vez instalado. El programa se instala con éxito en diferentes arquitecturas bajo los sistemas operativos Linux y Mac OS.

El software S3det se ha implementado así mismo en el servidor web Treedet (**Fig. 9**; Carro *et al.* 2006), accesible en la dirección http://treedetv2.bioinfo.cnio.es (**) con la colaboración de Ángel Carro y la Dra. Ana M. Rojas. Treedet incluye así los resultados de 3 métodos que representan los principales abordajes en la detección de residuos implicados en especificidad funcional (**Tabla 2 y 3**), esto es: el S-Method (uso de árboles; del Sol *et al.* 2003), el MB (basado en correlaciones; del Sol *et al.* 2003) y S3det (basado en análisis multivariante). Los tres métodos se acompañan de la herramienta Square (Tress *et al.* 2004) para la evaluación de la calidad del alineamiento múltiple de entrada. De este modo, Treedet proporciona una visión integrada de los posibles residuos dentro de un MSA con potencial interés funcional y permite una exploración sistemática del espacio de secuencias.

En la **Tabla 4** se muestran los resultados sobre diferentes tests del rendimiento de Treedet de forma desglosada. Puede observarse la alta eficiencia de S3det medida en tiempo de cálculo comparada con el resto de métodos para todos los casos ensayados.

Caso sencillo (MSA con 97 proteínas y 165 posiciones)					
Carga del test	**S3det**	**MB**	**S**	**Square**	**Total**
Test simple (una petición al servidor incluyendo todos los métodos)	2 min	3 min	7 min	5 min	22 min
Test complejo (quince peticiones simultáneas al servidor incluyendo todos los métodos)	2 min	5 min	7 min	20 min	3 h 44 min
Caso complejo (MSA con 198 proteínas y 412 posiciones)					
Carga del test	**S3det**	**MB**	**S**	**Square**	**Total**
Test simple (una petición al servidor incluyendo todos los métodos)	2 min	8 min	27 min	10 min	55 min
Test complejo (quince peticiones simultáneas al servidor incluyendo todos los métodos)	10 min	1h 57 min	3h 39 min	3h 30 min	6h 6 min

Tabla 4. Resultados de diferentes test de rendimiento del servidor web Treedet desglosado para los diferentes métodos que lo integran.

Consideradas en conjunto, las implementaciones de S3det permiten disponer de un software accesible, distribuíble y altamente eficiente que facilita su uso a la comunidad científica.

** Puede accederse al servidor desde todo tipo de navegadores excepto *Internet Explorer*.

Fig. 9. Captura de pantalla del servidor Treedet accesible en la dirección *web* http://treedetv2.bioinfo.cnio.es. Se observa la integración del software S3det junto a los métodos S, MB y Square (ver texto).

Parte II
Estudio a gran escala de la contribución de importantes aspectos de la especificidad funcional a la evolución en secuencia de las familias de proteínas

El desarrollo de S3det permite abordar a gran escala el estudio de la correspondencia de diferentes aspectos funcionales con subfamilias y SDPs de forma coherente, al estar estas definidas bajo un mismo marco metodológico (tal como se abordó en la Introducción (**Sección I.9**), la definición de subfamilias y SDPs es mutuamente dependiente). Se describe a continuación un estudio sobre un amplio conjunto de familias donde se analiza la contribución de importantes aspectos de la especificidad funcional a la evolución en secuencia de las familias de proteínas eucariotas. Este estudio se ha llevado a cabo en colaboración con David de Juan y el Dr. Florencio Pazos bajo la dirección del Prof. Alfonso Valencia y ha sido publicado en la revista *Proceedings of the National Academy of Science USA* (2010; la publicación se adjunta en el Anexo II).

En la **Fig. 10** se esquematiza la estrategia seguida en este estudio. Se eligieron dos características fundamentales en la función biológica de las proteínas: por un lado, el mecanismo catalítico (en el caso de los enzimas) y la unión específica a pequeños ligandos; por otro, las interacciones proteína-proteína. Su relación con las características en secuencia de las proteínas que componen la familia se estudia en base a dos entidades: las subfamilias en las que se organizan internamente y las posiciones del alineamiento múltiple asociadas a esta segregación (SDPs). Así, en el caso de las subfamilias de proteínas se analiza su correspondencia con etiquetas diferenciales que reflejan la especificidad funcional en los dos aspectos anteriormente citados: diferentes clases enzimáticas representadas por los códigos EC y diferentes conjuntos de interactores. Esto es, se quiere indagar si las diferentes subfamilias encontradas responden a funciones específicas diferentes. En el caso de las SDPs características de una familia de proteínas, se estudiará su implicación en los aspectos de la especificidad funcional citados a través de su asociación estructural a sitios de unión a ligando y sitios catalíticos así como a regiones de interacción.

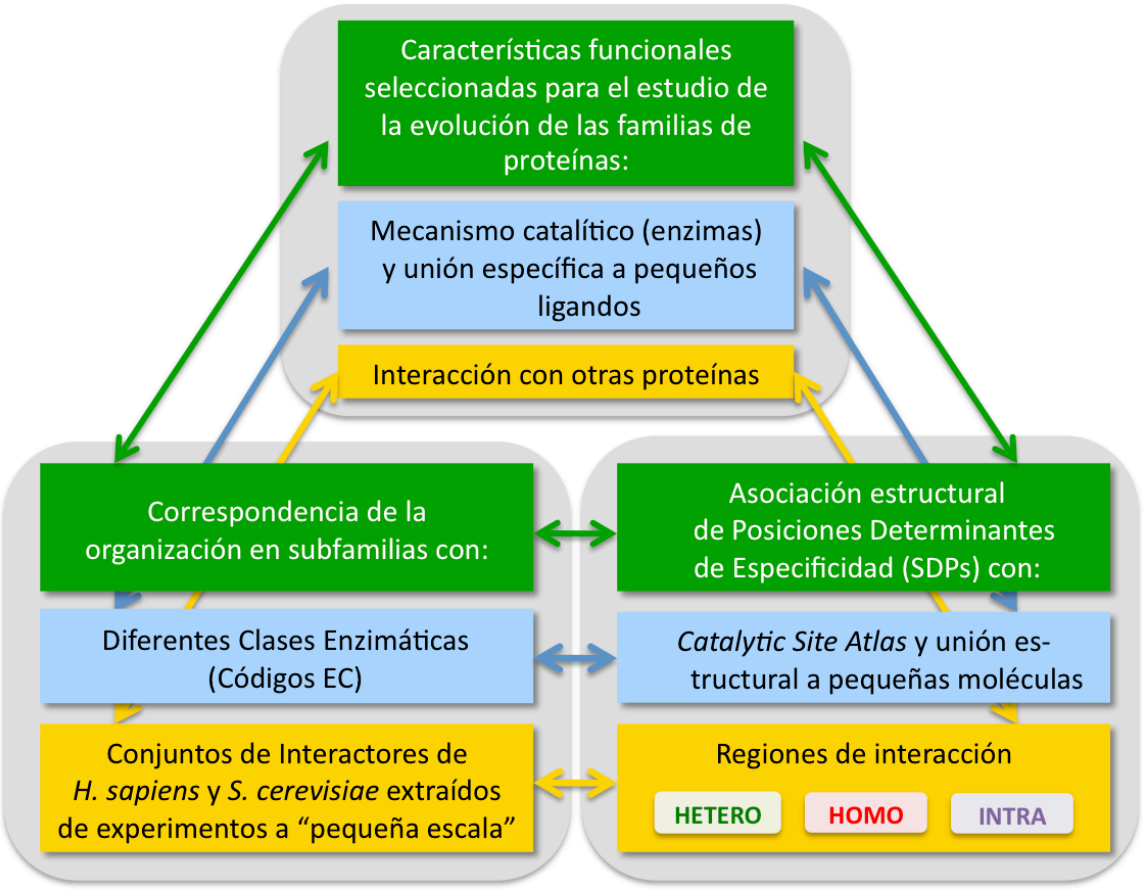


Fig. 10. Esquema de la estrategia seguida en el estudio a gran escala de la contribución de importantes aspectos de la especificidad funcional a la evolución en secuencia de las familias de proteínas.

Para llevar a cabo este estudio a gran escala se partió de un conjunto de 1262 familias Pfam (Bateman *et al.* 2004) de proteínas eucariotas (**Fig. 11**; **Métodos Sección VI.I**). La definición de subfamilias y SDPs en cada una de las familias Pfam se realizó bajo un mismo marco metodológico mediante la aplicación del método S3det anteriormente expuesto. Los resultados del análisis funcional de las subfamilias y SDPs obtenidos se detallan en las siguientes secciones.

a)

N° total de familias Pfam en el conjunto de partida	1262
Para el estudio de la correspondencia de características funcionales con la organización en subfamilias	
N° de familias con clasificación enzimática diferencial (códigos EC)	149
N° de familias con interactores diferenciales	72
Para el estudio de la correspondencia de SDPs con regiones estructurales de importancia funcional	
N° de familias con dominio estructural SCOP y sitio de unión a ligando o interfaz	316 (230)
N° de familias con sitio de unión a ligando e interfaz	168 (127)
N° de familias con sitio de unión a ligando	209 (152)
N° de familias con interfaz	276 (205)
N° de familias con interfaz Heteromérica	87 (73)
N° de familias con interfaz Homomérica	171 (137)
N° de familias con interfaz intra-cadena	170 (130)

** Los números entre paréntesis indican el número correspondiente de dominios SCOP no redundantes*

b)

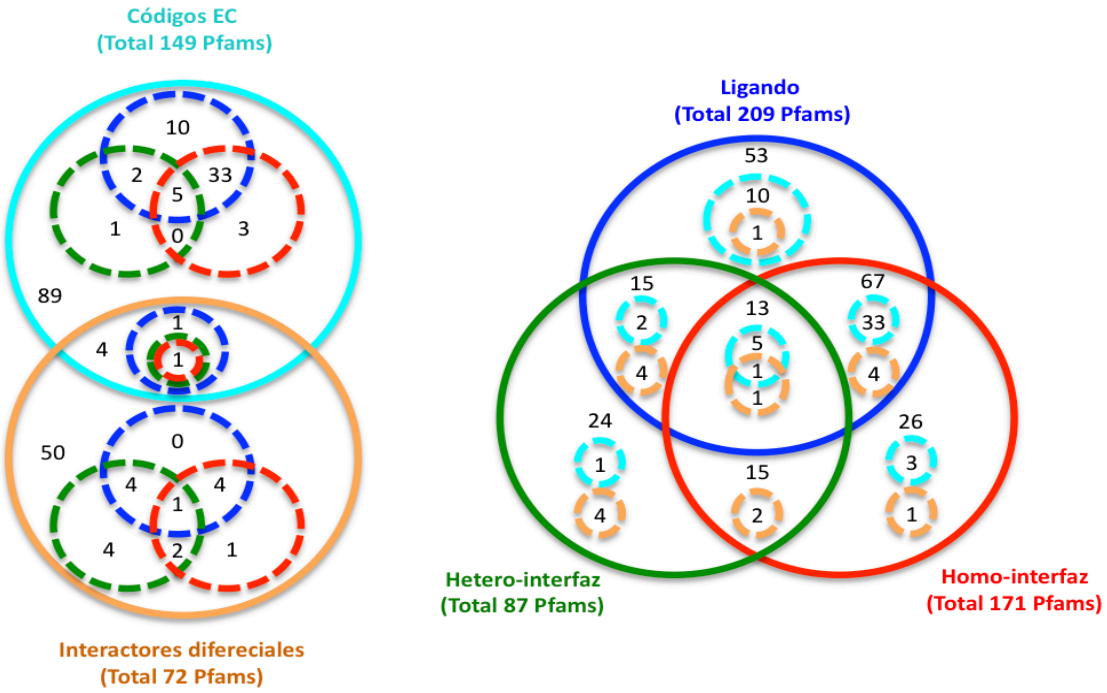


Fig. 11. Conjuntos de familias Pfam analizados en función de la información funcional y estructural recopilada. (a) Número de familias en los conjuntos estudiados. **(b)** Diagramas de Venn representado el solape entre algunos de los conjuntos señalados.

III.II.1. Estudio de la correspondencia entre subfamilias y etiquetas funcionales: Especificidad bioquímica y en las interacciónes proteína-proteína

Para el estudio de la correspondencia entre la organización en subfamilias y clases enzimáticas diferenciales se utilizó un conjunto de 149 familias Pfam para las cuales se dispone de suficiente información en términos de códigos EC (**Fig. 11**) y que superaron los filtros descritos en Métodos (**Sección VI.II.1**; en esta sección se detalla el procedimiento seguido para la obtención de esta información y la selección de familias con este tipo de diversidad funcional).

Para el análisis de la correspondencia de subfamilias con conjuntos diferentes de interactores se utilizó un conjunto de 72 familias Pfam para las cuales se obtuvo suficiente número de interacciones proteína-proteína, tanto positivas como negativas, en los organismos *S. cerevisiae* y *Homo sapiens* (**Métodos Sección VI.II.2**; **Fig. 11**). Como breve recordatorio, los pares de interacciones positivas se extrajeron de los experimentos etiquetados como "pequeña escala" en el núcleo (*core*) de la base de datos DIP (Xenarios *et al.*, 2000). El conjunto de pares negativos (pares de proteínas que se asume que no interaccionan) para cada uno de estos dos organismos se infirió a partir de los pares de proteínas que pertenecen a diferentes rutas metabólicas o que no comparten localización subcelular. Como requisito adicional en la definición de negativos, se exige siempre que el par de proteínas no se encuentre anotado como positivo en BIOGRID (Stark *et al.* 2006), repositorio general de conjuntos de datos de proteínas interactoras que incluye experimentos masivos.

El ajuste entre subfamilias y etiquetas funcionales (grupos de EC o interactores diferenciales) se calcula en este trabajo en términos de especificidad / sensibilidad a través de un análisis ROC. Para cada familia Pfam dentro del conjunto respectivo, el cálculo de la sensibilidad y la especificidad se realiza como sigue:

$$Sensibilidad = \frac{\sum TP}{\sum TP + \sum FN}$$

$$Especificidad = \frac{\sum TN}{\sum TN + \sum FP}$$

Donde:

- TP (del inglés *True Positives*) es el número de pares de proteínas que coinciden tanto en etiqueta funcional como en subfamilia.

- TN (del inglés *True Negatives*) es el número de pares de proteínas que no coinciden ni en etiqueta funcional ni en subfamilia
- FN (del inglés *False Negatives*): es el número de pares de proteínas que coinciden en su etiqueta funcional pero no en la subfamilia
- FP (del inglés *False Positives*): número de pares de proteínas que coinciden en subfamilia pero no en su etiqueta funcional.

En el análisis sólo se consideraron casos para los cuales se tiene que TP+FN>0, FP+TN>0, TP+FP>0 and FN+TN>0.

En el caso de los códigos EC, definir si dos proteínas comparten o no etiqueta funcional es trivial. En cambio, la situación se hace más compleja cuando se trata de evaluar el parecido de dos proteínas en base a su perfil de interactores y no interactores. Para este propósito se define aquí un Ratio de Interactores Comunes (SIR, del inglés *Shared Interactors Ratio*) para cada par de proteínas de una familia como sigue:

$$SIR = P^{++}/(P^{++}+P^{+-})$$

Donde:

- P^{++} representa el número de interactores común a ambas proteínas
- P^{+-} representa el número de interactores de una u otra proteína que no interaccionan con la otra.

Mientras que en el caso del código EC cada par de proteínas posee un valor binario (0/1) representando su coincidencia o no en esta etiqueta funcional, para el caso de las interacciones se tiene un valor de similitud continuo, desde 0.0 (no comparten interactores) hasta 1.0 (comparten todos sus interactores). Así, el cálculo de la sensibilidad y especificidad anterior se adapta para el tratamiento de esta similitud funcional "continua" como sigue:

- TP=∑(SIR) sobre los pares de proteínas pertenecientes a la misma subfamilia.
- TN=∑(1-SIR) sobre los pares de proteínas pertenecientes a diferentes subfamilias.
- FN=∑(SIR) sobre los pares de proteínas pertenecientes a diferentes subfamilias.
- FP=∑(1-SIR) sobre los pares de proteínas pertenecientes a la misma subfamilia.

Mediante el procedimiento explicado se logra resumir en un par de valores (especificidad y sensibilidad) el ajuste que, para una familia Pfam dada, se observa entre sus subfamilias y las etiquetas funcionales obtenidas (grupos de EC o interacciones diferenciales). Así por ejemplo, una familia Pfam con una especificidad de 1.0 implica que todas las proteínas dentro de una misma

subfamilia poseen la misma etiqueta funcional, y esto ocurre para cualquiera de sus subfamilias. Por otra parte, una familia Pfam con una sensibilidad de 1.0 implica que todas las proteínas con la misma etiqueta funcional recaen en la misma subfamilia, y esto ocurre para cualquiera de sus etiquetas funcionales. En el caso de la familia de aminotransferasas de clase III anteriormente expuesto (Pfam PF00202; **Fig. 4** izquierda; **Sección III.I.2.A**) los valores obtenidos -considerando sus códigos EC- fueron: sensibilidad=1 y especificidad=0.88. En el caso de la familia de factores de transcripción E2F/TDP (Pfam PF02319; **Sección III.I.2.B**) los valores obtenidos –considerando aquí sus interactores específicos (**Fig. 8b**)– fueron: sensibilidad=0.65 y especificidad=0.77.

Los valores de sensibilidad y especificidad obtenidos para cada una de las 149 familias con etiquetas EC y las 72 con interactores diferenciales se resumen en la **Fig. 12** a través de su representación en sendos espacios ROC (del inglés *Receiver Operating Characteristic*) donde en abcisas se muestra el valor "1-especificidad" y en ordenadas la sensibilidad. Una familia en la que exista una equivalencia perfecta entre sus subfamilias y sus etiquetas funcionales estaría representada en la esquina (0.0, 1.0) del área ROC. Los resultados en este gráfico se han representado en forma de histograma bidimensional en intervalos de 0.1 para ambas magnitudes. De esta manera, en cada región discreta se representa el porcentaje de familias con valores de sensibilidad y especificidad dentro del intervalo correspondiente. Los valores porcentuales se han redondeado al entero más cercano en aras de la simplicidad y su representación gráfica se acompaña mediante cuadrados coloreados cuyo lado es proporcional al valor correspondiente.

Los espacios ROC muestran una concordancia general entre la estructura en subfamilias y la distribución en ellas de ambos tipos de especificidad funcional, tanto en términos de ECs como en interactores específicos. La comparación detallada de los resultados obtenidos para ambas clasificaciones funcionales debe hacerse teniendo en cuenta sus respectivas naturalezas y los tamaños de grupo que implican, esto es: la fracción de proteínas dentro de una familia (alineamiento Pfam) que se sabe experimentalmente que interaccionan con la misma pareja (interactor) es generalmente pequeña comparada con la fracción de proteínas dentro de una familia que comparten código EC. Esto conduce por un lado a sensibilidades mayores en el caso de las agrupaciones de proteínas en base a sus interactores diferenciales (fila superior en el espacio ROC) y, por el otro, a especificidades superiores en el caso de las etiquetas EC. Es decir, mientras que las proteínas con el mismo EC tienden a incluir más de una subfamilia, las subfamilias tienden a incluir grupos de proteínas con más de un patrón de interacción diferencial.

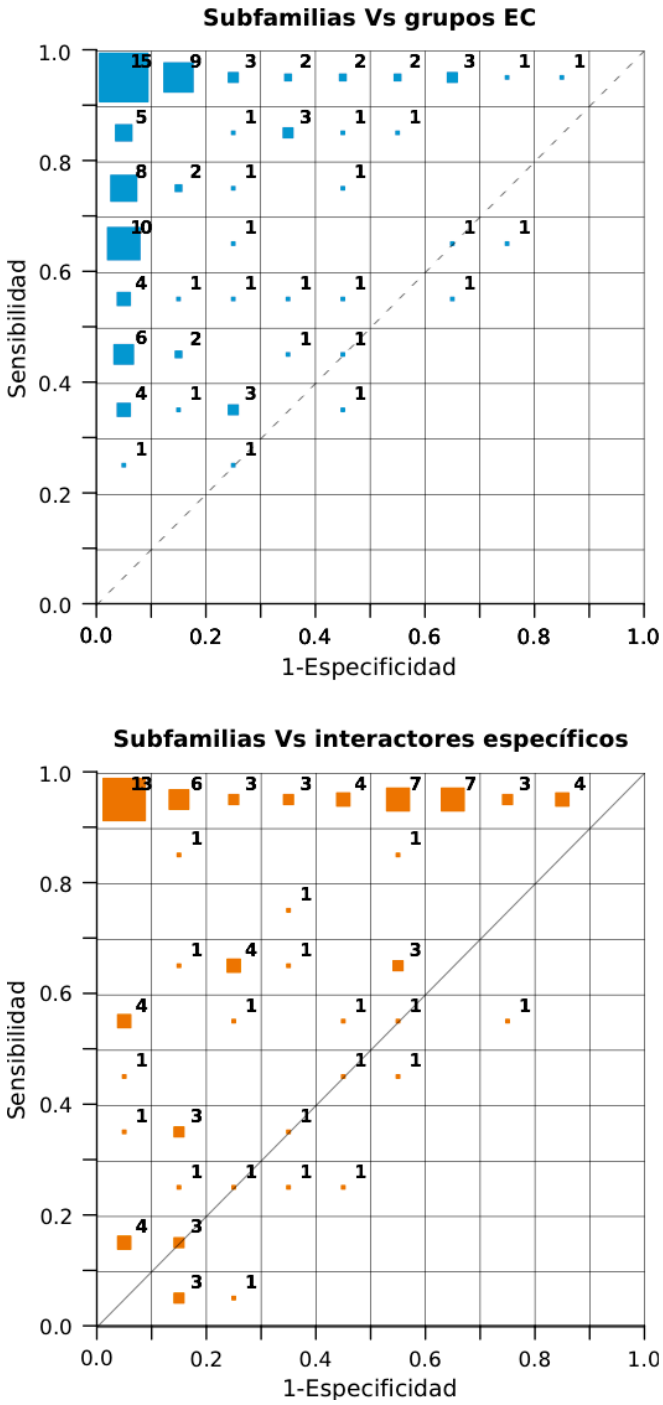


Fig. 12. Correspondencia entre las diferentes subfamilias y los grupos de ECs (arriba) e interactores específicos (abajo). Los valores de sensibilidad y especificidad obtenidos para cada una de las familias Pfam estudiadas se representan en sendos espacios ROC en los que la distribución de familias se muestra en forma de histograma bidimensional. En cada región discreta se representa el porcentaje de familias con valores de sensibilidad y sensibilidad dentro del intervalo correspondiente. Los valores porcentuales se han redondeado al entero más cercano en aras de la simplicidad (por lo que su suma puede ser diferente de cien) y su representación gráfica se acompaña mediante cuadrados coloreados cuyo lado es proporcional al valor correspondiente.

No obstante, se observa que la mayoría de familias presentan valores muy buenos de especificidad y sensibilidad, lo cual refleja una buena concordancia entre su estructura en subfamilias y su diversidad en especificidades funcionales. El hecho de que este resultado se observe simultáneamente para ambas etiquetas funcionales (ECs e interacciones específicas) indica que los patrones de interacción diferencial se encuentran integrados de forma coherente en la estructura de las familias de proteínas, y esto en un grado similar –de forma relativa- a la especificidad de su función bioquímica recogida por la clasificación EC, tal vez mejor caracterizada.

Este mismo análisis se realizó a partir de la definición subfamilias obtenida a través de métodos alternativos aplicados sobre el mismo conjunto de familias. Los resultados de estos métodos (**Anexo I, Fig. S1**) soportan y dan generalidad al acuerdo global de las subfamilias de proteínas tanto con grupos diferenciales de EC como con grupos de interactores específicos.

Para complementar estos análisis en términos de cobertura sobre el número de familias Pfam y usando otra definición de clase funcional, se realizó un estudio equivalente haciendo uso de la información funcional implícita en los identificadores de proteínas SwissProt ID, tal como se ha hecho en otros trabajos (p.ej. Dessimoz e*t al.* 2006). Así, se considera que dos proteínas pertenecen al mismo grupo "ID" si la primera parte de su identificador coincide (p.ej. OAT_HUMAN, OAT_MOUSE, OAT_YEAST, etc.). La primera parte del identificador SwissProt ID representa una abreviación del nombre del gen/proteína correspondiente, y se comparte entre secuencias que la base de datos SwissProt considera ortólogas (representando el organismo al que corresponden mediante la segunda parte del identificador).

De esta manera se obtuvo un conjunto de 799 familias Pfam para las cuales se dispone de suficiente información en términos de identificadores SwissProt IDs diferenciales (**Métodos Sección VI.II.3**). Para cada una de estas familias se calculó su correspondencia entre subfamilias e identificadores en términos de sensibilidad y especificidad de forma equivalente al caso de los códigos ECs presentado. Los resultados obtenidos se muestran en la **Fig. 13** donde, para un gran número de familias, se observa una consistencia general entre subfamilias y grupos de ortólogos funcionalmente homogéneos. Esta consistencia se aprecia fundamentalmente en los altos valores de sensibilidad observados, mostrando una tendencia en las organizaciones en subfamilias a reunir grupos de IDs sin escindirlos.

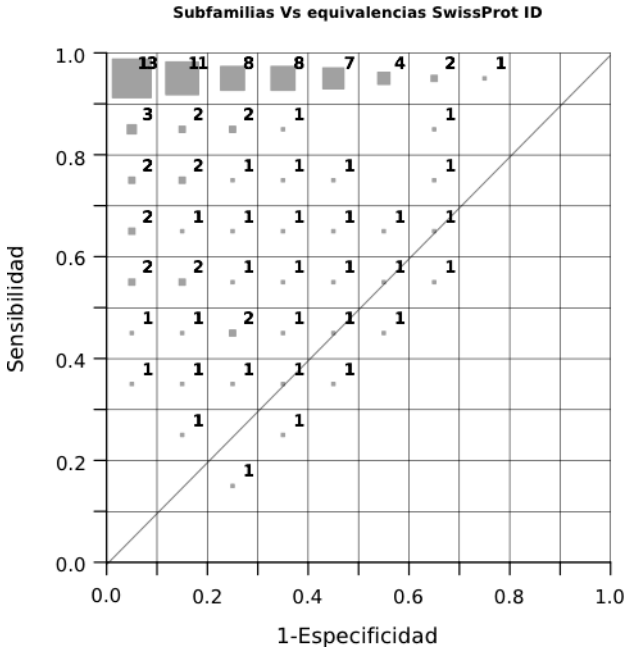


Fig. 13. Correspondencia entre las diferentes subfamilias y los grupos de identificadores de proteínas SwissProt ID. Los valores de sensibilidad y especificidad obtenidos para cada una de las 799 familias Pfam estudiadas se representan en un espacio ROC en el que la distribución de familias se muestra en forma de histograma bidimensional. En cada región discreta se representa el porcentaje de familias con valores de sensibilidad y especificidad dentro del intervalo correspondiente. Los valores porcentuales se han redondeado al entero más cercano en aras de la simplicidad (por lo que su suma puede ser diferente de cien) y su representación gráfica se realiza mediante cuadrados coloreados cuyo lado es proporcional al valor correspondiente.

III.II.2. Estudio de la asociación estructural entre SDPs y regiones funcionales: sitios de unión a ligando e interfaces.

La relación entre SDPs y regiones funcionales se investiga aquí en términos de su proximidad estructural a i) sitios de unión a ligando y sitios catalíticos, y ii) sitios de interacción con proteínas. Los primeros están conceptualmente asociados a las funciones bioquímicas descritas por el código EC analizadas previamente . De modo análogo, las regiones de interacción de proteínas están relacionadas con los patrones de interacción diferencial expuesto también en la sección anterior.

Para este propósito, a partir del conjunto inicial de 1262 familias Pfam, se recopiló el máximo posible de información fiable a nivel estructural a partir de proteínas y complejos con estructura resuelta. En **Métodos Sección VI.III** se explica en detalle el procedimiento seguido en este sentido. En esencia, para cada familia de proteínas se obtiene el conjunto de estructuras PDB disponibles, previo filtro por criterios de calidad de las estructuras y de la correspondencia del dominio Pfam con algún dominio estructural de SCOP. Para cada estructura PDB

se obtiene entonces, según el caso, el conjunto de residuos que forman parte del sitio de unión a ligando, que están anotados en el *Catalytic Site Atlas* (Porter *et al.*, 2004) y/o que forman parte de interfaces de interacción (pudiendo ser estas entre cadenas o entre el correspondiente dominio SCOP y el resto de la cadena). Una vez determinados los residuos funcionales para cada estructura del alineamiento Pfam, se elige un PDB como representante de toda la familia y sobre él se acumulan –a través del propio alineamiento Pfam- los residuos funcionales del resto, además de los suyos propios. Los criterios de calidad seguidos a fin de asegurar la máxima fiabilidad en la información estructural se detallan en las secciones de Métodos apuntadas.

Como resultado se obtuvieron 209 familias de dominios Pfam con sitios de unión a ligando conocido y 276 familias con regiones de interacción detectable (**Fig. 11**). Los SDPs de estas familias -obtenidos a través de S3det- así como el conjunto de posiciones conservadas de cada alineamiento (definidas al 90% de identidad de secuencia) se mapearon en las estructuras PDB representantes de cada alineamiento y sobre ellas se realizó el cálculo de las distancias a los residuos funcionales. La **Fig. 14** representa las distribuciones de distancias promediadas por familia entre SDPs, sitios de unión a ligando y regiones de interacción.

El análisis de estas distribuciones muestra que las SDPs están significativamente más próximas a las “regiones funcionales” (mediana 9.38 ± 5.28 Å -en el caso de los sitios de unión a ligando- y 7.60 ± 6.04 Å -en el caso de las interfaces) que el promedio del conjunto total de posiciones -fondo o *background*- con 11.82 ± 4.03 Å y 9.14 ± 4.76 Å respectivamente. En comparación, las posiciones conservadas (definidas al 90% de identidad en secuencia) se encuentran también próximas a las regiones funcionales (7.94 ± 4.40 Å y 7.16 ± 5.45 Å) y, en promedio, incluso más cercanas que el conjunto de SDPs. Estas diferencias están soportadas por los p-valores obtenidos a partir de un test de Wilcoxon para datos apareados, con valores inferiores a 1e-13 -en el caso de las SDPs- y <1e-15 -en el caso de las posiciones conservadas- cuando se evalúan contra los sitios de unión a ligando, mientras que los p-valores correspondientes a su evaluación contra superficies de interacción son <1e-9 -en el caso de las SDPs- y <1e-15 -en el caso de las posiciones conservadas (**Fig. 14**).

El desplazamiento de las SDPs y las posiciones conservadas hacia las regiones funcionales es significativo incluso teniendo en cuenta que el conocimiento sobre estas es incompleto: se sabe que los sitios anotados en PDB son sólo una parte de los sitios catalíticos y/o biológicamente activos, al igual que las interfaces descritas en estructuras 3D son sólo una pequeña parte de las interacciones reales entre proteínas. Otro factor a tener en cuenta es que, aquellas posiciones de un PDB próximas a un sitio de unión a ligando pero

lejanas a una interfaz se tratan como “negativos” en términos de su relevancia funcional cuando se evalúan exclusivamente sobre el segundo aspecto, y al inversa. En definitiva, estas consideraciones apuntan a la posibilidad de estar tomando como negativas muchas posiciones que podrían no serlo.

a)

b)

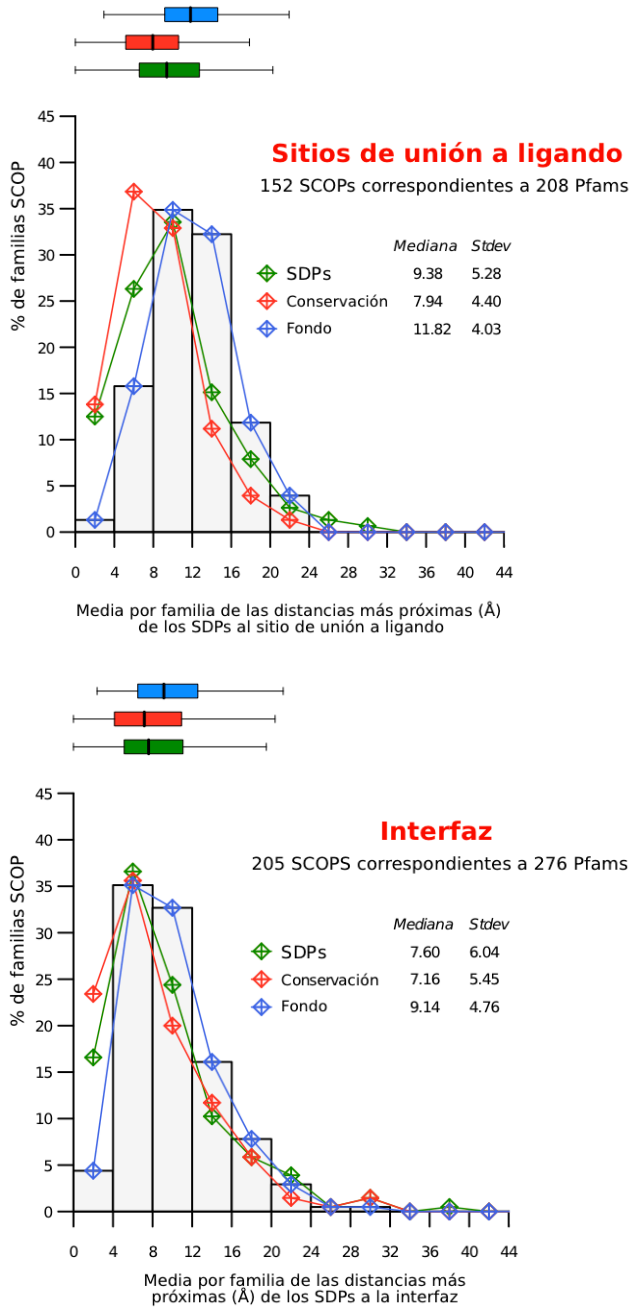


Figura 14. Distribución de las distancias mínimas entre las SDPs (en verde)**, las posiciones conservadas** (en rojo) **y el fondo** (en azul y en histogramas grises) **al sitio de unión a ligando** (arriba) **y a la interfaz** (abajo)**, promediadas dentro de cada familia Pfam y dentro de cada grupo estructuralmente redundante según SCOP**. Las distancias entre residuos

corresponden a las distancias entre sus átomos C_β–C_β. En la parte superior de ambos gráficos se resumen estas distribuciones mediante sus correspondientes diagramas de caja (*boxplots*) siguiendo el mismo código de colores.

Un análisis más exigente consiste en evaluar no ya si una determinada SDP se encuentra próxima a una región funcional sino si esta posición forma parte de hecho de la región misma (anotada como sitio de unión a ligando o a proteínas). Con este propósito, se evalúa aquí -a través de tests de Wilcoxon de suma de rangos- si existe o no un enriquecimiento de residuos funcionales en el conjunto de SDPs y para un número significativo de familias por encima de lo que es esperable al azar (**Fig. 15**; **Métodos Sección VI.V**).

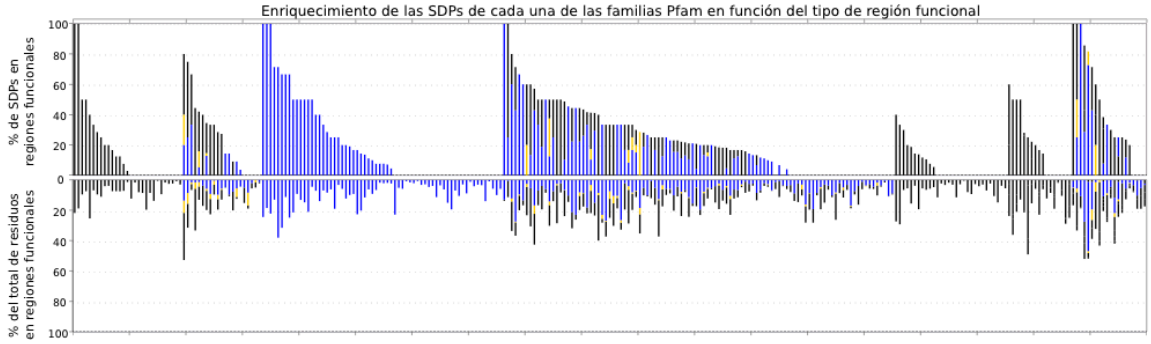


Fig. 15. Porcentaje de SDPs en regiones funcionales (arriba) **comparado con el porcentaje que esas mismas regiones funcionales representan sobre el total de residuos de la proteína** (fondo o *background;* abajo) **para cada familia Pfam** (cada barra corresponde a una familia). Los sitios de unión a ligando se representan en azul, las interfaces proteína-proteína en negro y el solape entre ambas regiones en amarillo. Las interfaces intra-cadena se han omitido en la figura en aras de la simplicidad.

Los p-valores obtenidos (**Tabla 5**) muestran que tanto las SDPs como las posiciones conservadas se encuentran claramente enriquecidas en: i) sitios anotados de unión a ligando, ii) sitios de unión a proteína (interfaces) y iii) la combinación de ambos tipos de posiciones funcionales tomadas como un todo. Puede también observarse que los enriquecimientos son generalmente más significativos respecto a los sitios de unión a ligando que a los de unión a proteínas. Todos estos tests se han realizado asumiendo que sólo se dispone de información de secuencia, ya que tanto las SDPs como las posiciones conservadas se extraen de MSAs. Si la evaluación se restringe a las posiciones en la superficie de la proteína, los enriquecimientos anteriores -tanto para SDPs como para la conservación- aumentan su significación de forma apreciable (**Tabla 5**).

Los mismos test de enriquecimiento funcional se aplicaron a las SDPs obtenidas a través de métodos alternativos para el mismo conjunto de familias

(**Anexo I**, **Tabla S1**). Los resultados de estos análisis (**Anexo I, Tabla S2**) soportan y dan generalidad a la asociación de SPDs tanto con sitios de unión a ligando como con regiones de interacción.

p-valor	Total residuos funcionales	Sitio de unión a ligando	Interfaz total
SDPs *(sobre el total)*	**1.67E-05**	**4.25E-04**	**1.89E-02**
SDPs *(restringido a superficie)*	**1.00E-07**	**2.79E-04**	**4.02E-04**
Conservación *(sobre el total)*	**1.92E-27**	**4.73E-20**	**1.92E-09**
Conservación *(restringido a superficie)*	**2.65E-30**	**1.21E-18**	**2.54E-15**

Diferencia de medianas	Total residuos funcionales	Sitio de unión a ligando	Interfaz total
SDPs *(sobre el total)*	**5.32%**	**3.97%**	**2%**
SDPs *(restringido a superficie)*	**8.69%**	**5.53%**	**4.28%**
Conservación *(sobre el total)*	**15.25%**	**14.60%**	**6.34%**
Conservación *(restringido a superficie)*	**22.73%**	**18.66%**	**11.71%**

Tabla 5. Resultados de los test de Wilcoxon de suma de rangos evaluando el enriquecimiento de SDPs y posiciones conservadas en sitios funcionales.

III.II.3. La asociación funcional de SDPs en proteínas que poseen tanto sitios de unión a ligando como regiones de interacción con proteínas.

Una vez establecida estadísticamente la asociación de SDPs con sitios de unión a ligando e interfaces, se aborda la cuestión de si existe una implicación preferente de las SDPs entre éstos dos tipos de regiones funcionales. Se restringe por tanto aquí el análisis a las 168 familias de proteínas para las cuales se conoce ambos tipos de regiones (**Fig. 11**). Sobre ellas se analiza la distribución de distancias de SDPs a sitios de unión a ligando y a interfaces de forma conjunta, a través de sus pares de distancias a una y otra región. Para desentrañar esta asociación, se pone el énfasis en la cantidad total de

posiciones de las 168 familias que se encuentran: i) próximas a una región pero lejanas de la otra región; ii) a la inversa; iii) cercanas a ambas; o iv) lejanas a ambas simultáneamente. Este análisis complementa, por lo demás, el abordaje implícito en los anteriores tests presentados en la sección anterior, en los cuales la tendencia a la proximidad se cuantifica de forma relativa, corrigiendo por el tamaño de las regiones en cuestión (las interfaces son por lo general de mayor tamaño que los sitios de unión a ligando).

En la **figura 16a** se muestran los resultados de este análisis. Se observa que existen posiciones (tanto SDPs como conservadas) cercanas a la interfaz pero lejanas del sitio de unión a ligando, y viceversa. Si se establece sobre el gráfico una distancia de contacto de 8 Å entre carbonos beta, se pueden establecer cuatro regímenes: a) posiciones muy cercanas (en contacto) tanto con sitios de unión a ligando como a interfaces (~23% del total de SDPs; ~26 % del total de posiciones conservadas); posiciones en contacto con el sitio de unión a ligando pero no con la interfaz (~17% SDPs; ~21% posiciones conservadas); posiciones en contacto con la interfaz pero no con el sitio de unión a ligando (~28% SDPs; ~25% posiciones conservadas); y finalmente posiciones que no contactan a ninguno de los dos tipos de regiones funcionales (~32% SDPs; ~25% posiciones conservadas). En la **figura 16b** se muestra una distribución equivalente en la que las distancias han sido promediadas por familia Pfam y por grupo estructuralmente redundante según SCOP. Estas distribuciones muestran que la implicación relativa de SDPs en sitios de unión a ligando e interfaces es similar en términos de distancias.

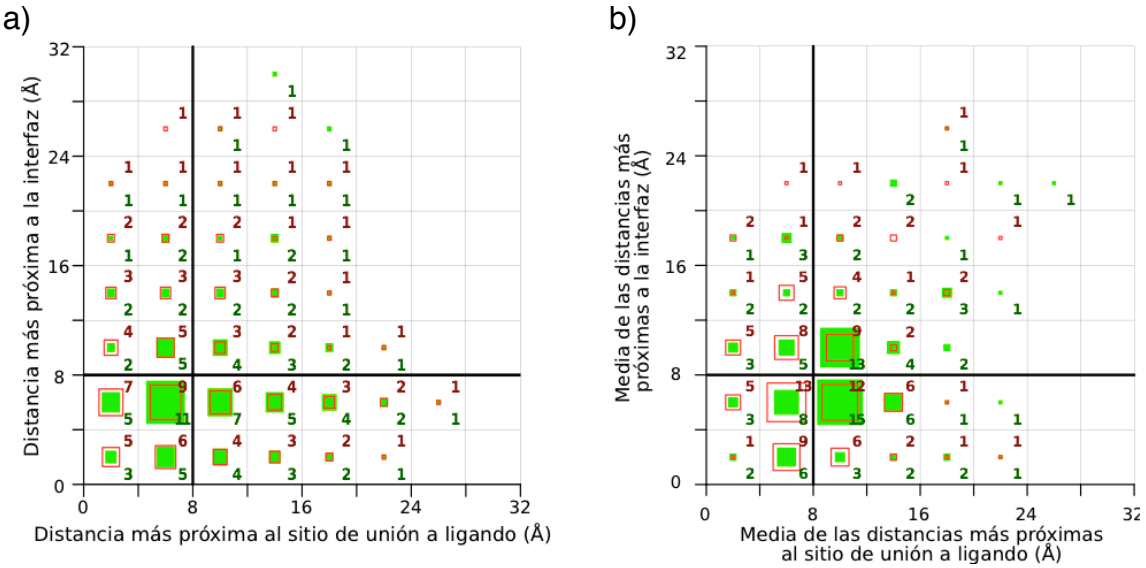


Fig. 16. Histogramas bidimensionales representando la distribución conjunta de las distancias más cercanas entre SDPs (en verde) **y posiciones conservadas** (en rojo) **al sitio de unión a ligando** (en abcisas) **y la interfaz** (en ordenadas)**. (A)** Distribución conjunta de distancias del total de SDPs de las 168 familias Pfam que presentan ambos tipos de regiones funcionales. El lado de los cuadrados es proporcional al porcentaje de SDPs y posiciones

conservadas que contiene cada intervalo (calculados sobre el total de posiciones de cada tipo –SDPs y conservadas- por separado), el valor del cual se indica también en forma de número. **(B)** Distribución conjunta análoga a la anterior en la que las distancias han sido promediadas por familia Pfam y por grupo estructuralmente redundante de acuerdo a SCOP. El lado de los cuadros es proporcional al porcentaje -sobre el total de grupos- de grupos de familias no redundantes a nivel estructural en cada intervalo del histograma, el valor del cual se indica también en forma de número. En aras de la simplicidad, los valores porcentuales se han redondeado al entero más cercano por lo que su suma puede ser diferente de cien.

Para complementar este último análisis, se calcula aquí el número de familias Pfam en las cuales al menos uno de sus SDPs forma parte de hecho de: i) el sitio de unión a ligando; ii) la interfaz; o iii) ambos tipos de regiones funcionales (en el caso de que estas sean solapantes). Se trata pues de forma cualitativa el/los tipo/s de región funcional responsable/s de la especificidad funcional dentro de una familia a partir de la naturaleza de las regiones funcionales en las que mapean sus SDPs. En la **Tabla 6** se resumen estos cálculos para todas las combinaciones posibles.

Familias con al menos una de sus SDPs en el **sitio de unión a ligando** (excluyendo la región de solape, si la hay)	Familias con al menos una de sus SDPs en la **intersección** entre ambas regiones funcionales	Familias con al menos una de sus SDPs en la **interfaz** (excluyendo la región de solape, si la hay)			
+	**+**	**+**	9.1%	19.4%	41.2%
+	**+**	**-**	1.8%		
-	**+**	**+**	4.8%		
-	**+**	**-**	3.6%		
+	**-**	**+**	21.8%		
+	**-**	**-**	15.8%		
-	**-**	**+**	23.6%		
-	**-**	**-**	19.4%		
48.5%	19.4%	59.4%			
57.0%					
	64.8%				
80.6%					

Tabla 6. Porcentajes condicionales acumulados de familias en las que al menos uno de sus SDPs forma parte del sitio de unión a ligando, la interfaz, o el solape entre ambas regiones. Los porcentajes se calculan sobre el total de 168 familias Pfam para las cuales existe evidencia estructural de ambos tipos de regiones funcionales**.**

De las 168 familias de proteínas que presentan ambos tipos de regiones funcionales, el 57.0% tienen alguna SDP que forma parte del sitio de unión a ligando, el 64.8% de las familias tienen alguna SDP que forma parte de la interfaz y el 41.2% tienen SDPs en ambos tipos. De los porcentajes condicionales acumulados obtenidos (**Tabla 6**) no puede derivarse que en el

conjunto de familias estudiadas exista una clara tendencia en cuanto al tipo de región funcional en el que mapean las SDPs.

III.II.4. Estudio desglosado de la asociación estructural entre SDPs y distintos tipos de regiones de interacción: hetero-, homo- e intra-interfaces.

Una vez demostrada la asociación entre SDPs y regiones de interacción para un número estadísticamente significativo de familias, se pretende aquí indagar en mayor detalle la naturaleza de esta asociación desglosándola en tres tipos de interfaces: las que median la interacción entre proteínas diferentes (heterocomplejos), las que implican dos proteínas idénticas (homocomplejos) y las que se definen en base a la interacción entre el dominio en estudio y el resto de la cadena a la que pertenece (intra-cadena). Así, el conjunto de familias para las cuales se obtuvo información estructural de regiones de interacción se escindió de acuerdo a estos tres tipos de interfaz obteniendo tres conjuntos de 87, 171 y 170 familias respectivamente (**Fig. 11**). Consecuentemente, en cada uno de estos conjuntos sólo se tuvo en cuenta ese tipo de interacciones en la definición de los residuos que forman parte de la interfaz. En la **Fig. 17** se muestran los resultados obtenidos para las familias Pfam en función de diferentes tipos de interfaz.

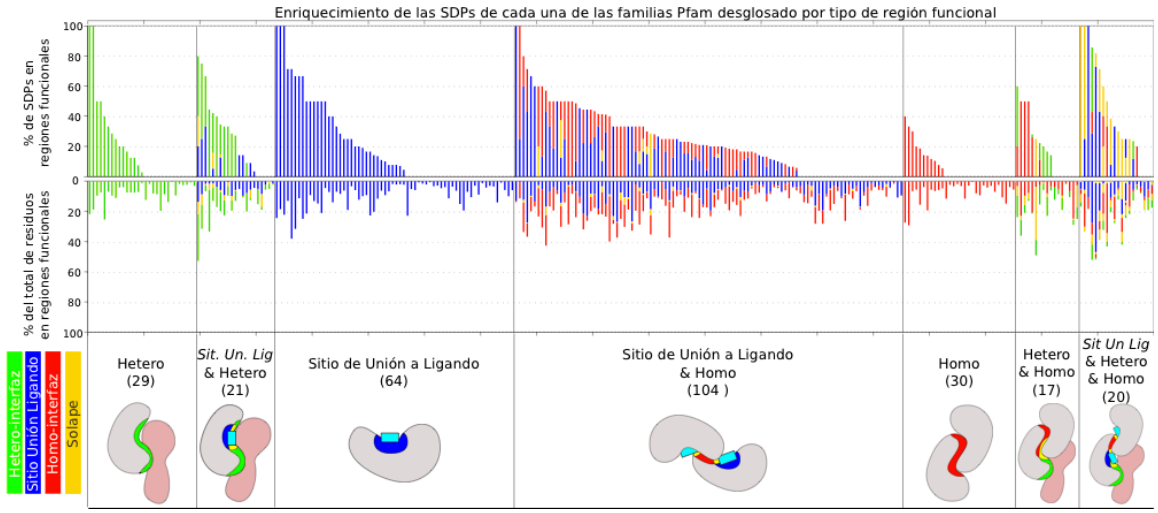


Fig. 17. Porcentaje de SDPs en regiones funcionales (desglosando interfaces homo- y hetero-) (arriba) **comparado con el porcentaje que esas mismas regiones funcionales representan sobre el total de residuos de la proteína (**fondo o *background;* en el medio**) para cada familia Pfam** (cada barra corresponde a una familia). Las familias se agrupan en función del tipo de región funcional detectada (abajo), y el número de familias en cada grupo se muestra entre paréntesis. Los sitios de unión a ligando se representan en azul, las interfaces heterodiméricas en verde, las homodiméricas en rojo y sus solapes en amarillo. Las interfaces intra-cadena se han omitido en aras de la simplicidad.

Se realizaron a continuación los test de enriquecimiento de SDPs y posiciones conservadas en cada uno de estos tres conjuntos, de forma equivalente a los ya mostrados en la **Sección III.III.** anterior. Los resultados de estos test se muestran en la **Tabla 7**. Se observa que las SDPs están significativamente enriquecidas en interfaces que implican heterocomplejos (p-valor<0.05). Sin embargo, el enriquecimiento no se da en un número significativo de casos para el resto de interfaces (homodiméricas o intracadena) cuando estas se consideran de forma aislada. Las posiciones conservadas sí se encuentran en cambio estadísticamente asociadas a los tres tipos de interfaz. Cuando los test se restringen a las posiciones en superficie (tanto para SDPs y posiciones conservadas como para el fondo), se obtiene el mismo tipo de enriquecimientos significativos, si bien por lo general con mejores p-valores.

p-valor	Total residuos funcionales	Sitio de unión a ligando	Interfaz total	HETERO	HOMO	INTRA
SDPs *(sobre el total)*	**1.67E-05**	**4.25E-04**	**1.89E-02**	**1.75E-02**	5.13E-01	4.99E-01
SDPs *(restringido a superficie)*	**1.00E-07**	**2.79E-04**	**4.02E-04**	**1.56E-02**	8.75E-01	1.13E-01
Conservación *(sobre el total)*	**1.92E-27**	**4.73E-20**	**1.92E-09**	**5.32E-02**	**9.09E-03**	**4.29E-07**
Conservación *(restringido a superficie)*	**2.65E-30**	**1.21E-18**	**2.54E-15**	**1.19E-03**	**1.27E-04**	**1.60E-10**

Diferencia de medianas	Total residuos funcionales	Sitio de unión a ligando	Interfaz total	HETERO	HOMO	INTRA
SDPs *(sobre el total)*	**5.32%**	**3.97%**	**2.00%**	**4.77%**	-0.01%	0.00%
SDPs *(restringido a superficie)*	**8.69%**	**5.53%**	**4.28%**	**7.40%**	1.32%	1.30%
Conservación *(sobre el total)*	**15.25%**	**14.60%**	**6.34%**	**1.84%**	**1.99%**	**6.76%**
Conservación *(restringido a superficie)*	**22.73%**	**18.66%**	**11.71%**	**5.60%**	**4.30%**	**11.27%**

Tabla 7. Resultados de los test de Wilcoxon de suma de rangos evaluando el enriquecimiento de SDPs y posiciones conservadas en diferentes tipos de interfaz. Se incluyen los mismos resultados que en la Tabla 5 a fin de favorecer la comparación de los valores con los enriquecimientos obtenidos para las regiones funcionales consideradas de forma agregada.

De nuevo, estos resultados son consistentes con los obtenidos para estos mismos tests a partir de SDPs obtenidos mediante métodos alternativos (**Anexo I, Tablas S1 y S2**).

Para desglosar con mayor potencia estadística la implicación de SDPs entre los diferentes tipos de interfaz se requeriría de un conjunto mayor de casos en los que se den al mismo tiempo interfaces de tipo *hetero*, *homo* e *intra*. Sin embargo, la intersección de los conjuntos de familias obtenidos para cada una de ellas es relativamente baja. P.ej. la intersección entre los conjuntos *homo* y *hetero* se reduce a 37 familias (**Fig. 11b**), observándose por lo demás un grado variable de solape entre ambos tipos de interfaz (**Fig. 17**).

Parte III
Desarrollo y aplicación de métodos supervisados para la predicción de posiciones determinantes de especificidad funcional

Se desarrollaron dos nuevos métodos supervisados para la predicción de posiciones determinantes de especificidad funcional: Xdet y MCdet. Estos métodos predicen residuos funcionales a partir de alineamientos múltiples de proteínas (alineamientos que pueden estar basados en estructura o de secuencia). Se califican como supervisados porque explotan información funcional conocida, que puede ser tanto de tipo cuantitativo como en forma de clasificación cualitativa. Los métodos Xdet y MCdet se aplican en esta sección al estudio de alineamientos de proteínas en los que las relaciones de similitud en secuencia no se corresponden con sus parecidos relativos respecto al aspecto funcional estudiado.

El desarrollo de estos métodos y las aplicaciones que se describen en esta sección se realizaron en colaboración con el Dr. Florencio Pazos bajo la dirección del Prof. Alfonso Valencia y fueron publicados por la revista *Bioinformatics* (2006; la publicación se adjunta en el **Anexo II** para su consulta).

III.III.1. El método Xdet

Este método tiene como objetivo detectar las posiciones de un alineamiento múltiple de proteínas responsables de su clasificación funcional cuando entre las diferentes especificidades se puede establecer una relación de similitud jerárquica o cuantificar una distancia. La idea que explota este abordaje es que, en este tipo de posiciones, un cambio brusco en el tipo de aminoácido entre dos proteínas estaría relacionado con una diferencia funcional sustancial, y a la inversa.

En la **Fig. 18** se muestra un esquema explicativo del método Xdet. Para cada posición del alineamiento se construye una matriz cuantificando la similitud entre los tipos de aminoácidos de cada par de proteínas (parecido que puede extraerse, por ejemplo, de una matriz estándar de sustitución). En esta matriz, un elemento dado representa el parecido entre los residuos de dos proteínas en esa posición. Por otra parte, se construye una matriz -a partir de información funcional conocida- en la que cada elemento representa el parecido funcional entre el correspondiente par de proteínas (para la característica funcional concreta en la que estemos interesados). Las dos matrices así construidas se comparan con el coeficiente de correlación de rango de Spearman "r_k" (Press *et al.* 1992; ver más abajo en esta sección).

Las posiciones con elevado r_k serán aquellas en las que se de una alta correlación entre los parecidos entre aminoácidos y los parecidos funcionales y, por tanto, son las posiciones que se predecirán como responsables de esos tipos de especificidad funcional.

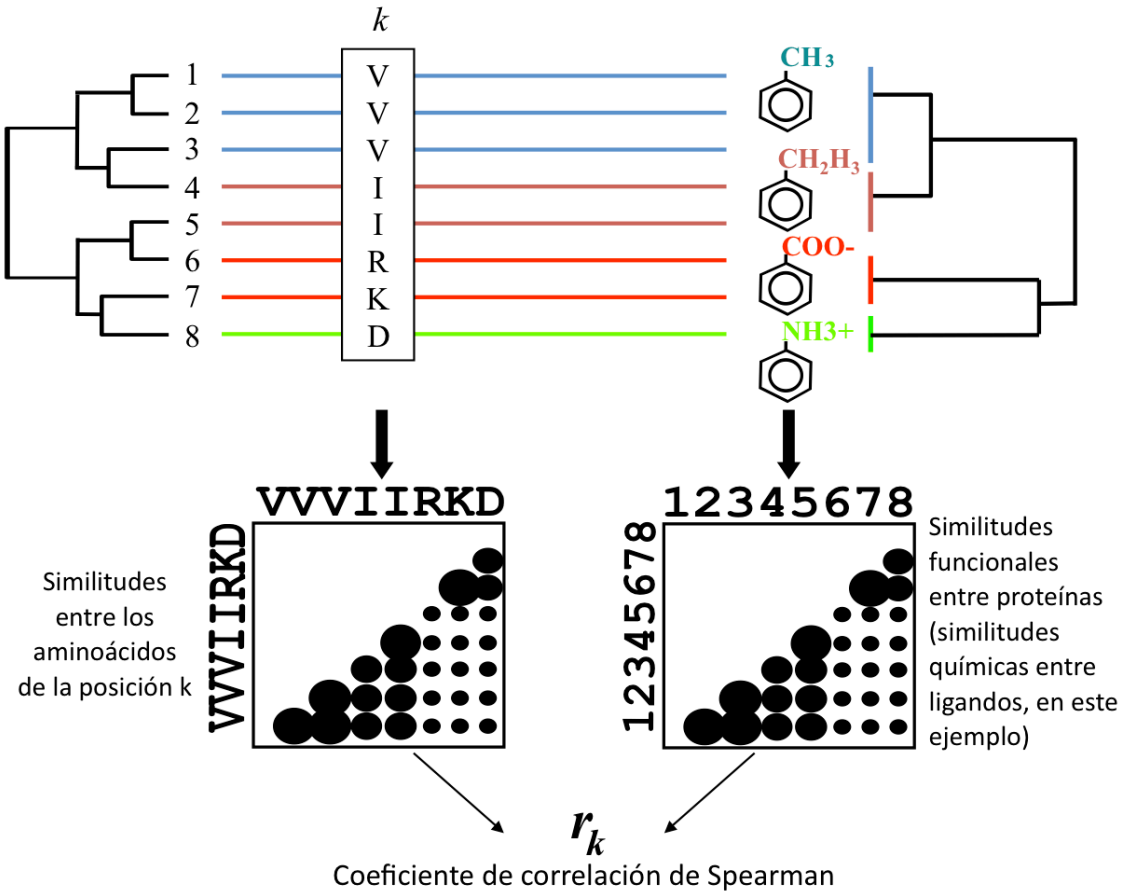


Fig. 18. Esquema explicativo del método Xdet. Se representa un alineamiento esquemático de ocho secuencias de proteínas. Las relaciones filogenéticas implícitas entre las secuencias se muestra a su izquierda en forma de árbol. Los colores de las proteínas representan una clasificación funcional que, en el supuesto representado, no refleja las relaciones de similitud en secuencia implícitas en el alineamiento. En el esquema, la función (representada a la derecha) es la unión a una molécula pequeña, con ligeras variaciones químicas uniendo a diferentes proteínas de la familia. En este caso se pueden derivar similitudes funcionales entre las proteínas a partir de las similitudes químicas de los ligandos a los que unen. Las similitudes funcionales pueden a su vez representarse en forma de matriz o de árbol (clasificación jerárquica), tal como se representa en la parte derecha de la figura. Para calcular si una determinada posición *k* del alineamiento está relacionada con la jerarquía funcional, se construye una matriz de similitud análoga en la que se evalúa el parecido entre los tipos de aminoácidos que aparecen en esa posición. Los valores de estas matrices se representan en la figura mediante círculos de radio proporcional al parecido entre los pares comparados. La comparación global entre ambas matrices de similitud se realiza mediante el coeficiente de correlación de rango de Spearman (Press *et al.* 1992).

Como medida de similitud funcional entre proteínas se pueden utilizar diversos tipos de métricas, dependiendo del problema que estemos tratando así como la definición asociada de función. Así por ejemplo puede utilizarse el parecido químico entre ligandos, métricas para medir el parecido en clasificaciones funcionales jerárquicas al estilo de las implícitas en *Gene Ontology* (Harris *et al.*, 2004) o el código enzimático (EC), jerarquías funcionales basadas en conocimiento experto, el parecido entre parámetros enzimáticos (K_{cat},K_m,…), y otras . En el caso extremo en el que la clasificación funcional no permita derivar similitudes funcionales de forma cuantitativa, se puede asignar un valor binario (0/1) para representar la semejanza entre pares de proteínas que pertenecen a la misma o distinta clase funcional, respectivamente.

Se dedica el resto de esta sección a los detalles formales de la implementación del método Xdet:

Dado un MSA con *N* proteínas y *P* posiciones, se define A_{ijk} como la semejanza entre los aminoácidos de las proteínas *i* y *j* en la posición *k*, y F_{ij} como la similitud funcional entre las proteínas *i* y *j*. El coeficiente de correlación de rango de Spearman "r_k" se calcula para cada posición *k* como sigue (**Fig. 18**):

$$r_k = \frac{\sum_{i,j} \left[(A'_{ijk} - \overline{A'}) \cdot (F'_{ij} - \overline{F'}) \right]}{\sqrt{\sum_{i,j} (A'_{ijk} - \overline{A'})^2} \cdot \sqrt{\sum_{i,j} (F'_{ij} - \overline{F'})^2}} \qquad (18)$$

donde *A'* y *F'* son los valores de rango de *A* y *F* respectivamente (en los casos de empate, se asignan los rangos medios; Press *et al.*, 1992). $\overline{A'}$ y $\overline{F'}$ son los correspondientes valores medios de rango de las respectivas matrices. Las posiciones con más de un 10% de huecos (*gaps*) se excluyen de los cálculos anteriores y como valor de similitud entre un hueco y cualesquiera de los aminoácidos se asigna el valor nulo ("0").

Las posiciones con elevado r_k serán aquellas en las que se dé una alta correlación entre los parecidos entre aminoácidos y los parecidos funcionales y, por tanto, son las posiciones que se predecirán como responsables de esos tipos de especificidad funcional.

Para estos resultados se asignan p-valores mediante su comparación con una distribución aleatoria de r_k obtenida a partir de 1000 asignaciones función-proteína al azar sobre el mismo MSA y las mismas funciones.

III.III.2. El método MCdet

El método MCdet está directamente relacionado con el método S3det presentado en esta tesis (**Resultados Sección III.I.1**). Si bien se desarrolló con anterioridad a S3det, MCdet puede considerarse metodológicamente una simplificación de aquel y sus aspectos matemáticos son parcialmente solapantes. No obstante, se describe aquí de forma completa a fin de que pueda accederse de forma independiente.

Este método parte de una representación vectorial de secuencias, residuos y funciones. Esta representación permite el estudio de la relación entre estos tres conjuntos de elementos mediante técnicas de análisis vectorial, en este caso a través del Análisis de Correspondencias Múltiples (MCA). (Greenacre, 1984; Lebart *et al.*, 1984).

En la **Fig. 19** se muestra un esquema explicativo del método MCdet. En primer lugar, el MSA se codifica como una matriz binaria (**Fig. 19, centro**; **Sección III.I.1.A**) que implica ya una primera representación vectorial tanto de proteínas como de residuos. El MCA aplicado sobre esta matriz realiza una transformación del sistema de coordenadas obteniendo los llamados "ejes principales" (**Fig. 19, abajo**). En estos ejes se produce una representación simultánea de secuencias y residuos en dos espacios íntimamente ligados y superponibles. La principal característica de los espacios obtenidos es que sus ejes son ortogonales. Por otra parte, permiten establecer una relación analítica entre ambos conjuntos de elementos (proteínas y residuos) a través de sus relaciones pseudovaricéntricas. De acuerdo a estas relaciones, el centro de masas de cualquier grupo de proteínas señala aquellos residuos particularmente asociados a ellas (**sección III.I.1.E**).

El tratamiento del MSA con MCA permite, por un lado, identificar grupos de proteínas particularmente vinculadas a determinados grupos de residuos (**Fig. 19, abajo)**. Bajo un escenario no supervisado, estas agrupaciones serían consideradas como las subfamilias de proteínas implícitas en el alineamiento y sus posiciones asociadas como las "Posiciones Determinantes de Especificidad" (SDPs) (Esta posibilidad es la que explota el método S3det que se presenta en la 1ª parte de estas tesis).

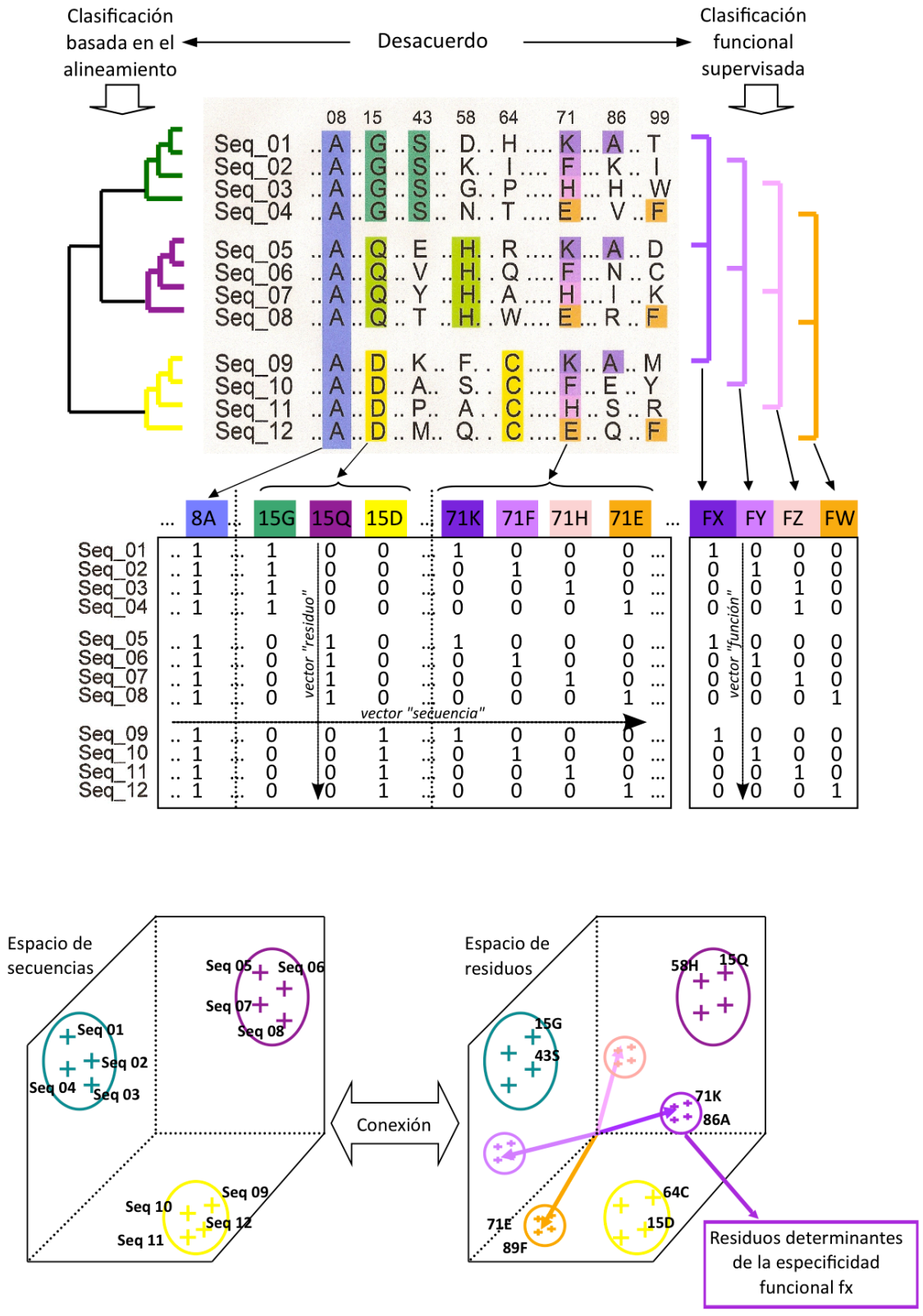


Fig. 19. Esquema explicativo del método MCdet. (Arriba) Se representa un alineamiento esquemático de doce secuencias de proteínas. Las relaciones de similitud entre las secuencias se muestran a su izquierda en forma de árbol. La clasificación funcional supervisada que se

quiere estudiar se muestra a la derecha del MSA a través de 4 funciones (*Fx*, *Fy*, *Fz* y *Fw*) mediante llaves conectoras que señalan las proteínas que pertenecen a cada una de las cuatro funciones. Se observa en este supuesto particular que la clasificación funcional no refleja las relaciones de parecido en secuencia correspondientes. **(Centro)** Codificación del MSA y de la clasificación funcional supervisada en forma de matrices binarias (Sección III.I.1.A). Estas matrices constituyen ya una primera representación vectorial tanto de proteínas y residuos como de funciones. **(Abajo)** Representación esquemática de los espacios principales generados a través del tratamiento con MCA del MSA codificado de forma binaria. En el espacio de secuencias, cada proteína del alineamiento original se representa mediante un punto. Se observa que las relaciones entre las distancias relativas de las proteínas reflejan sus distancias en secuencia. Los círculos en este espacio reflejan agrupaciones de proteínas que se asemejan más entre ellas que con el resto de proteínas. Bajo un escenario no supervisado (en el método S3det) estas agrupaciones se considerarían las subfamilias implícitas en el MSA. En el espacio de residuos, cada residuo en el alineamiento original se representa mediante un punto y cada función -codificada de forma binaria- se representa mediante un vector. Se observa que las funciones así representadas apuntan a conjuntos de residuos que caen en sus cercanías. Estos residuos son los que el método MCdet predice como responsables de determinar estas especificidades funcionales. En el mismo espacio aparecen otros conjuntos de residuos que siguen en cambio la distribución espacial equivalente a la de la segregación de las proteínas en el espacio de secuencias. En un abordaje no supervisado, estos últimos conjuntos de residuos serían predichos como los determinantes de la segregación en subfamilias y por tanto de sus atributos funcionales específicos (SDPs). Esta idea es la que explota el método S3det presentado en la 1ª parte de esta tesis.

Por otro lado, el MCA permite además proyectar -en el espacio de residuos- clasificaciones funcionales supervisadas que codifiquen de forma binaria la pertenencia o no de las proteínas del MSA a esa función (**Fig. 19; parte derecha**). Esta posibilidad es la que se ha implementado en el método MCdet de modo que, con independencia de cuál sea la segregación de proteínas en el espacio de secuencias, podemos identificar qué residuos del alineamiento se hallan particularmente asociados a una determinada función sencillamente con evaluar las distancias entre los “vectores residuo” y los “vectores función” (**Fig. 19, abajo**). De forma intuitiva, cuanto menor sea esta distancia, mejor será el ajuste entre el perfil de presencia/ausencia de una determinada función y el perfil de presencia/ausencia de un residuo dado. Los residuos con la menor distancia al vector que representa la función sobreimpuesta serán predichos como aquellos residuos que determinan sus características específicas.

Se dedica el resto de esta sección a los detalles formales de la implementación del método MCdet:

Dado un MSA con N secuencias y L posiciones, se construye una matriz W de dimensiones NxQ (donde $Q = 21 \cdot L$) en la que se representa cada posición l en el alineamiento como una categoría disjunta y completa con 21 modalidades (representando los 20 tipos de aminoácidos más la opción de hueco o *gap*)

codificando la presencia de una modalidad dada con un "1" y su ausencia con un "0". Las columnas en *W* sin ningún "1" son eliminadas en aras a la consistencia subsiguiente sin pérdida de generalidad, resultando en una matriz *X* de dimensiones *NxP*, donde *P<Q*. Dada la matriz X así obtenida, se definen las siguientes frecuencias como sigue:

$$x_{nS} = \sum_{P} x_{np} \qquad x_{Sp} = \sum_{N} x_{np} \qquad x_{SS} = \sum_{N}\sum_{P} x_{np} \tag{19}$$

$$f_{nS}=x_{nS}/x_{SS} \qquad f_{Sp}=x_{Sp}/x_{SS} \qquad f_{np}=x_{np}/x_{SS} \tag{20}$$

Sea *Y* la matriz de término general $y_{np}= f_{np}/(f_{Sp}\sqrt{f_{nS}})$. Considerar distancias euclídeas entre los vectores columna de *Y* equivale a considerar distancias Chi-cuadrado en la matriz de datos originales *X*. Sea *Z* la matriz de término general $z_{np}= f_{np} /\sqrt{(f_{nS}\cdot f_{Sp})}$ y Z^T su transpuesta. El espacio generado por los vectores principales de ZZ^T proporciona una descomposición de la asociación entre secuencias y residuos entre sus fuentes de variación (Peña, 2002).

A continuación, las columnas de la matriz Y se proyectan en el espacio generado por los vectores principales de ZZ^T. Sea v_k el *k*-avo vector principal correspondiente al *k*-avo valor propio λ_k no nulo de ZZ^T (excluyendo la solución trivial $\lambda=1$). Las coordenadas c_{pk} del residuo *p* en el eje principal *k* del espacio de secuencias vienen dadas por la expresión siguiente:

$$c_{pk} = \sum_{N} \frac{v_{kn}\, f_{np}}{f_{Sp}\sqrt{f_{nS}}} \tag{21}$$

Sea $\Phi=(\phi_1, \phi_2,\ldots..,\phi_N)$ un vector columna binario *Nx1* que representa una función Φ dada de tal manera que su término general ϕ_n es "1" si la secuencia *n* tiene la función *Φ*, ó "0" en caso contrario. En MCA, el vector *Φ* se puede proyectar directamente como una columna suplementaria en el espacio anteriormente generado de forma conjunta con los residuos "p". Este "vector función" mapeará en la proximidad de aquellos residuos cuyo patrón de ausencia/presencia a lo largo de la población de secuencias en el MSA se le parezca. Estos residuos se predecirán como aquellos residuos responsables de conferir las características distintivas de la función *Φ*.

Las coordenadas $c_{\Phi k}$ de la función *Φ* en el eje principal *k* del espacio de residuos vienen dadas por la expresión:

$$c_{\Phi k} = \frac{1}{\sum_{N}\Phi_n} \sum_{N}\left(\Phi_n \frac{v_{kn}}{\sqrt{f_{nS}}}\right) \tag{22}$$

Así pues, los residuos candidatos a ser los responsables de conferir la función *Φ* se determinan por aquellos *p* con una menor distancia euclídea a *Φ*, esto es:

$$d(p,\Phi) = \sqrt{\sum_{K} c_{pk}\, c_{\Phi k}} \qquad (23)$$

donde *K* es el máximo número de valores propios no nulos de ZZ^T. En el cálculo de las distancias *d(p,Φ)*, se consideran en este método el total de vectores principales *k*, los cuales equivalen a una varianza explicada del 100% (en esta aplicación particular del MCA se prescinde del filtrado de ejes principales; **Sección III.I.1.F**).

Al valor de la distancia *d(p,Φ)* se asigna un p-valor mediante su comparación con una distribución de distancias obtenida a partir de 1000 asignaciones aleatorias función-proteína para *Φ* sobre el mismo MSA.

III.III.3. Aplicación de los métodos Xdet y MCdet

Los métodos supervisados desarrollados se aplicaron a diferentes conjuntos de proteínas para las cuales se comprobó cuidadosamente que las características funcionales estudiadas no se correspondieran con sus parecidos relativos en secuencia. Los casos seleccionados ilustran diferentes escenarios donde se requiere el uso de métodos supervisados. Con ellos se cubre un amplio rango de situaciones en las que hay diferentes solapes entre la clasificación funcional y la basada en secuencia, diferentes formas de definir la función de las proteínas implicadas así como diferentes formas de cuantificar los parecidos funcionales. A continuación se presentan los conjuntos de proteínas seleccionados y los resultados obtenidos.

III.III.3.A. Homólogos estructurales del oncogén Ras

En este ejemplo se estudia la especificidad de unión a ligando de un conjunto de 24 proteínas estructuralmente homólogas al oncogén Ras (H-Ras-1 / *Transforming protein* p21 ; PDB 1ctqA). El alineamiento estructural del que se parte (**Métodos Sección VI.VI.I**) contiene proteínas de unión a diferentes tipos de ligando, incluyendo nucleótidos (GTP, GDP, FMN, FAD ...), nucleósidos, azúcares, etc.

En este caso, la ausencia de una correspondencia clara entre las similitudes en secuencia de las proteínas del alineamiento y la clasificación funcional (según el ligando al que se unen) se debe principalmente a que el alineamiento estructural (sumado a la eliminación de redundancia realizada,

Métodos Sección VI.VI.I) pone en relación homólogos remotos (cuyas distancias en secuencia son muy grandes). En la **figura 18a** se observa esta discrepancia a través del árbol obtenido para la familia (**Métodos Sección VI.VII**). Incluso en los grupos para los cuales hay un acuerdo de forma general entre su clasificación funcional y su segregación en el árbol (como p.ej. los GxP: GTP, GDP, GNP, etc; ver nota al pie †), hay excepciones notables como la proteína FtsZ de división celular (PDB 1fsz), que es una GTPasa lejana al grupo GxP.

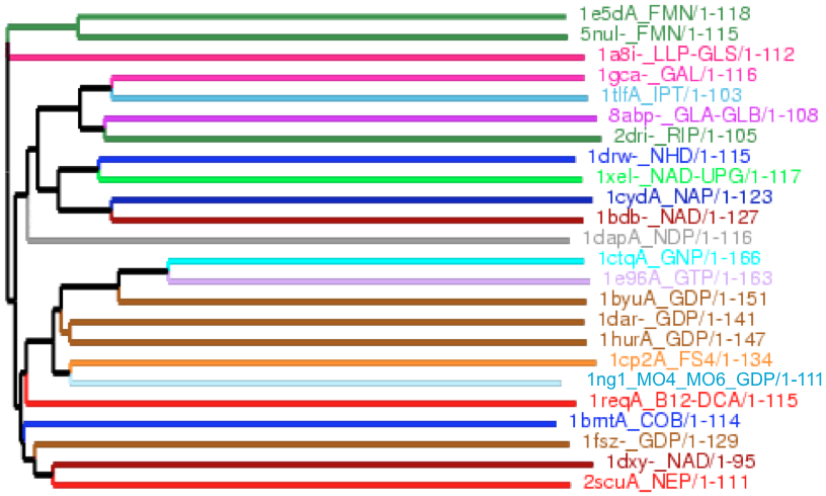
Este ejemplo permite también ilustrar los casos en que la clasificación funcional se deriva a partir del tipo de ligando unido a la proteína, y donde las similitudes funcionales entre las proteínas pueden cuantificarse a partir de las similitudes químicas entre sus respectivos ligandos (en este caso a través del coeficiente de Tanimoto, Holliday *et al.*, 2002)

El método MCdet se utilizó en este caso para predecir los residuos responsables de la función “unión a nucleótidos de guanina” (GxP) dado que es la única función con suficientes representantes en el alineamiento. Xdet, en cambio, está diseñado para predecir las posiciones con una importancia "global" en la determinación de las especificidades de unión, en lugar de estar focalizado sobre un tipo concreto de ligando.

En la **figura 18b** se muestran los residuos predichos por ambos métodos como responsables de la especificidad funcional de los homólogos estructurales del oncogén Ras, proyectados sobre la estructura de una proteína de unión a GTP, la RhoA de humanos (PDB 1ftn). Se representan sobre esta estructura los 11 residuos más cercanos al “vector función GxP” (ver nota al pie) según MCdet y los 6 residuos con mayor correlación respecto a la matriz de similitud funcional (3 de los cuales son comunes a ambos métodos).

† Todos los nucleótidos de guanina (GTP, GDP y GNP) se consideran como una sola clase funcional (referida como "GxP") dado que las propiedades de unión de las proteínas a las que unen son idénticas y el hecho de que se encuentren unidas a un nucleótido de guanina en particular se debe a factores extrínsecos a la proteína (p.ej. su cristalización con nucleótidos artificiales no hidrolizables, etc.). Así, dos proteínas que unan GTP y GNP respectivamente, son consideradas aquí funcionalmente idénticas.

a)



b)

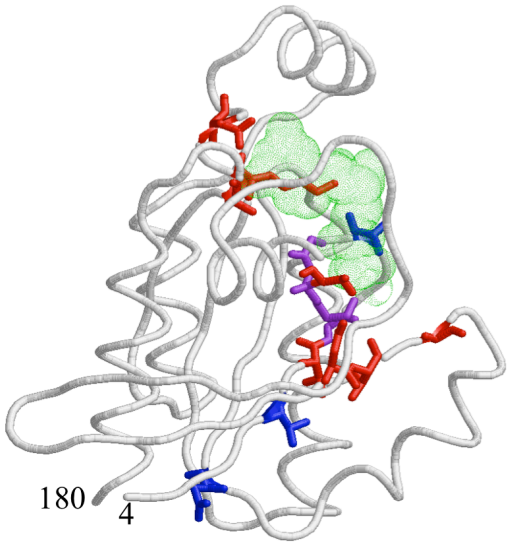


Fig. 20. Resultados de los métodos Xdet y MCdet para la familia de homólogos estructurales del oncogén Ras. (a) Árbol derivado del alineamiento estructural de este conjunto de proteínas obtenido por el método de *neighbour-joining* (Métodos Sección VI.VII). Las proteínas están etiquetadas de acuerdo con el código PDB de su estructura seguido del ligando al que se encuentran unidas (en nomenclatura PDB). Las ramas del árbol están coloreadas de acuerdo a los diferentes tipos de ligandos. Dos proteínas que unan a un determinado ligando pueden estar coloreadas de forma diferente si la lista completa de ligandos a los que unen no es exactamente la misma. **(b)** Residuos predichos por los métodos Xdet y MCdet mapeados sobre la estructura de la proteína RhoA de humanos (PDB 1ftn) de unión a GTP. En la estructura la proteína se encuentra unida a GDP, el cual se representa en forma de esferas color verde. Los residuos predichos por los métodos se representan mediante bastones (*sticks*) de color azul para los del método Xdet, rojo para los del método MCdet y violeta para las posiciones predichas por ambos métodos. La figura se generó mediante el programa Rasmol (Sayle y Milner-White 1995).

Las predicciones de ambos métodos se agrupan de forma evidente alrededor del nucleótido (GDP en esta estructura). Los residuos predichos por

MCdet recaen principalmente en la proximidad de la guanina y los grupos fosfato. Todas las predicciones de Xdet cercanas al ligando apuntan a los grupos fosfato, lo cual refleja que es esta la región donde se confiere especificidad "global" de unión a ligando en esta familia (la región donde los ligandos son más diferentes). De forma interesante, las predicciones de ambos métodos se alejan un poco más allá del último fosfato (*β*) del GDP, en la región donde el tercer fosfato (*γ*) se encontraría si la proteína se hubiese cristalizado en unión a GTP. Algunas otras predicciones (V9 y D78 para Xdet ó T60 y G62 para MCdet) mapean muy lejos del GDP. Para estas posiciones no se dispone de una interpretación evidente, pudiendo ser tal vez falsos positivos.

III.III.3.B. Dominios SH3

Este segundo ejemplo es un caso donde las proteínas relacionadas presentan homología remota en secuencia y su clasificación funcional se basa en conocimiento experto. Se pretende ilustrar cómo una clasificación funcional basada en ontologías jerárquicas puede utilizarse para cuantificar similitudes funcionales.

Los dominios SH3 son módulos de reconocimiento de péptidos que se unen a motivos ricos en prolina de ciertas proteínas (Zarrinpar *et al.* 2003). A pesar de tener un origen evolutivo común y una estructura general similar, estos dominios son muy divergentes en secuencia. Estos dominios son el caso prototípico de homología remota donde los métodos de predicción de residuos funcionales basados en secuencia son difíciles de aplicar.

Los dominios SH3 se pueden agrupar en diferentes clases funcionales dependiendo de los motivos característicos de los péptidos a los que se unen (Cesareni *et al.*, 2002; **Fig. 21a**). La similitud entre las clases mencionadas se derivó en este caso a partir de una clasificación funcional jerárquica de los dominios SH3 desarrollada por expertos (Cesareni *et al.* 2002, y comunicación personal). La **Figura 21b** muestra el árbol generado con el método unión de vecinos (*neighbour-joining*) a partir del alineamiento de los dominios SH3 (**Métodos Sección VI.VI.2**). Se observa que el árbol basado en este alineamiento no refleja la clasificación funcional asociada.

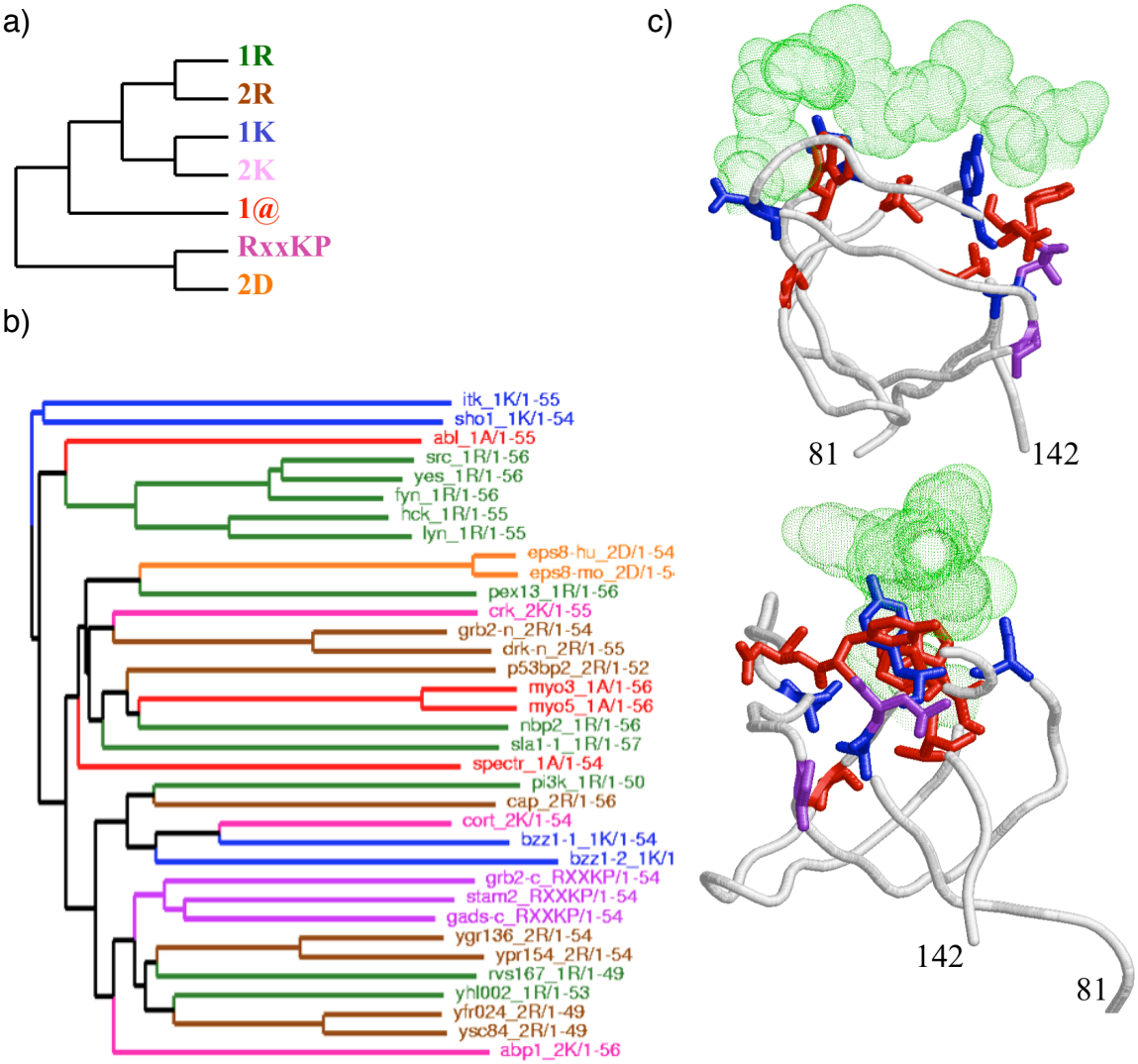


Fig. 21. Resultados de los métodos Xdet y MCdet para la familia de dominos SH3. (a) Clases funcionales de los dominios SH3 y sus correspondientes relaciones jerárquicas (adaptado de Cesareni *et al.* 2002). **(b)** Árbol obtenido por unión de vecinos (*neighbour joining*) a partir del alineamiento de los dominios SH3 (**Métodos Sección VI.VII**). Las ramas del árbol y sus correspondientes proteínas están coloreadas de acuerdo a las diferentes clases funcionales a las que pertenecen. **(c)** Se muestran dos vistas ortogonales de los residuos predichos por ambos métodos mapeados sobre la estructura de la tirosín-quinasa Fyn (PDB 1fyn). En la estructura, la proteína se encuentra unida a un péptido sintético el cual se representa en forma de esferas de color verde. Los residuos predichos por los métodos se representan mediante bastones (*sticks*) de color azul para los del método Xdet, rojo para los del método MCdet y violeta para las posiciones predichas por ambos métodos. La figura se generó mediante el programa *Rasmol* (Sayle y Milner-White 1995).

La **figura 21c** muestra las predicciones de ambos métodos mapeadas sobre la estructura del dominio SH3 de la tirosín-quinasa Fyn (perteneciente a la

clase “1R”) unida a un péptido sintético (PDB 1fyn). El método MCdet se utilizó para la predicción de los residuos responsables de la especificidad funcional "1R". Los conjuntos de residuos predichos por los dos métodos siguen claramente al péptido unido, concentrándose especialmente en sus extremos terminales. Se sabe que la especificidad de unión de los ligandos SH3 se determina principalmente sobre la base de la variabilidad en estos extremos mientras que la parte central del sitio de unión está más conservada a lo largo de las proteínas que unen ligandos diferentes (Cesareni *et al.* 2002). Las predicciones se extienden un poco más allá del extremo C-terminal del péptido. De hecho, las dos predicciones comunes entre ambos métodos se encuentran en esta región. Esto podría indicar que esta zona es también importante para determinar la unión con sustratos naturales, p.ej. proteínas de mayor tamaño. Algunos de los residuos predichos, como Y137 (puntuación más alta en Xdet) y L90, han sido ampliamente descritos en la literatura como importantes determinantes de la especificidad de unión a ligando (Cesareni *et al.* 2002).

III.III.3.C. Hidrolasas glicosídicas con estructura de barril TIM

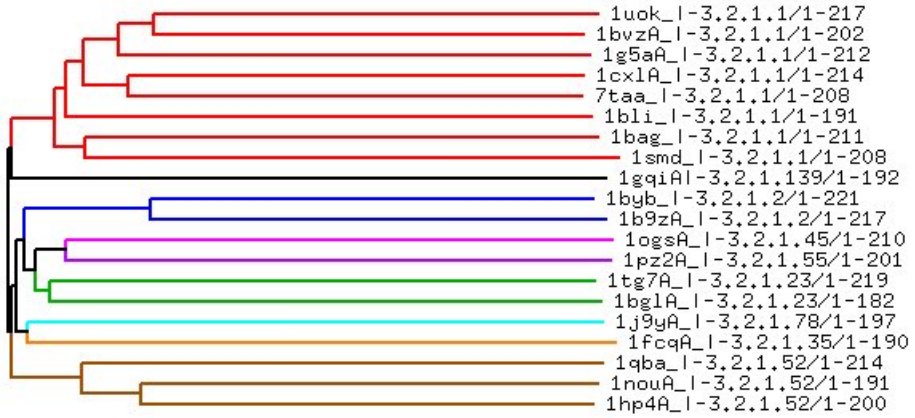
Se estudia aquí el alineamiento estructural de 20 hidrolasas glicosídicas (correspondientes al código enzimático EC 3.2.1.*) cuyas estructuras son barriles TIM (en inglés *TIM-barrels*; ver nota al pie[‡]). En este caso las clases funcionales vienen dadas por las actividades enzimáticas que difieren en su especificidad de sustrato (tomado aquí como la diferencia en su último número del código EC, compartiendo los 3 primeros dígitos). El alineamiento contiene secuencias pertenecientes a un total de 9 subclases diferentes de hidrolasas glicosídicas (3.2.1.1, 3.2.1.2, 3.2.1.35, ...).

El árbol obtenido a partir del alineamiento estructural sí refleja de forma general la separación de las 9 clases enzimáticas (**Fig. 22a**). No obstante, debido a la homología remota entre las proteínas implicadas, las distancias entre cada par de secuencias son extremadamente grandes, haciéndolas poco fiables para un análisis no supervisado basado en los parecidos relativos en secuencia (come se verá más adelante en esta sección).

El método Xdet se utilizó para predecir las posiciones con una importancia global en la diferenciación de las distintas clases dentro del código EC 3.2.1.*. En este caso, se puede realizar una cuantificación de las distancias funcionales de forma binaria, esto es: si un par de proteínas pertenecen a las misma clase se les asigna una similitud igual a 1, y 0 en el caso contrario. El método MCdet se utilizó para predecir las posiciones responsables de la especificidad 3.2.1.1 (correspondiente a la clase alfa-amilasa) por contraste con el resto de especificidades.

[‡] El barril TIM es un motivo estructural de proteínas muy conservado que consiste en ocho hélices alfa y ocho láminas beta paralelas que se alternan en el esqueleto de la proteína.

a)



b)

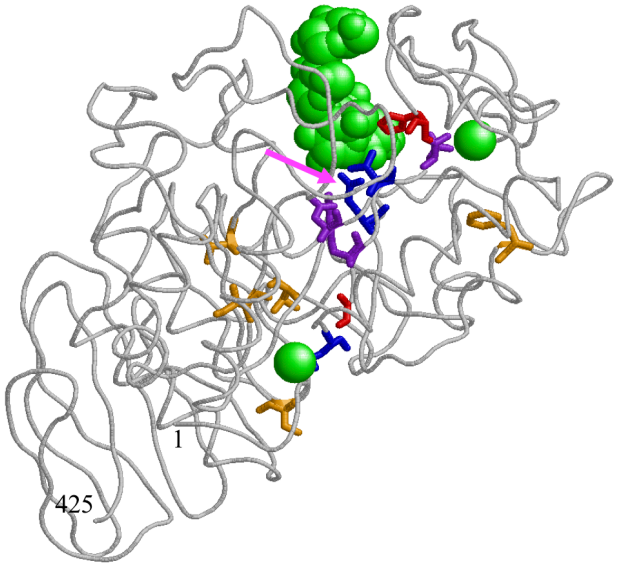


Fig. 22. Resultados de los métodos Xdet y MCdet para la familia de hidrolasas glicosídicas con estructura de barril TIM. (a) Árbol obtenido por unión de vecinos (*neighbour joining*) a partir del alineamiento estructural de glicosidasas de barril TIM descrito en Métodos Sección VI.VI.3. Las ramas del árbol y sus correspondientes proteínas están coloreadas de acuerdo a las diferentes clases funcionales a las que pertenecen. **(b)** Residuos predichos por los métodos Xdet y MCdet representados sobre la estructura de la alfa-amilasa de *B. subtilis* (PDB 1bag). En la estructura, la proteína se encuentra unida a maltopentosa y dos iones calcio, representados ambos en color verde. Los residuos predichos por los métodos se representan mediante bastones (*sticks*) de color azul para los del método *Xdet*, rojo para los del método MCdet y violeta para las posiciones predichas por ambos métodos. La flecha rosa indica la posición 208 mutada en Fujimoto *et al.* 1998. Los residuos predichos por el método MB (del Sol Mesa *et al.* 2003) -basado en la filogenia implícita en el alineamiento- se representan mediante bastones (*sticks*) de color naranja, a excepción del D97 que es también una predicción tanto de Xdet como de MCdet y se colorea por tanto en violeta. La figura se generó mediante el programa *Rasmol* (Sayle y Milner-White 1995)

La **figura 22b** muestra los residuos predichos por cada método mapeados sobre la estructura de la alfa-amilasa de *Bacillus subtilis* (PDB 1bag) (Fujimoto *et al.* 1998). Hay 3 residuos en común entre las 5 predicciones de MCdet y las 6 de Xdet. Hay por tanto un total de 8 residuos diferentes en el conjunto de predicciones de ambos métodos. Seis de estos ocho residuos se agrupan en el sitio activo de la proteína, indicado en este caso por la maltopentosa unida. El residuo con la segunda mejor puntuación arrojado por Xdet (el 208), se mutó (E208Q) por Fujimoto *et al.* (1998; el artículo original donde se presenta la estructura tridimensional 1bag) demostrando su participación en el sitio activo de la proteína. Dos de los residuos (D171 y G172) predichos de forma independiente por Xdet y MCdet (respectivamente) se encuentran relativamente lejos del sitio activo, aunque próximos uno al otro. El residuo D171 está coordinando a un ión calcio, lo cual podría apuntar a un posible rol funcional (si bien esta posibilidad no se menciona en la publicación original de Fujimoto *et al.*1998).

Para ilustrar la diferencia entre los métodos supervisados y los no supervisados, se aplicó el método MB (del Sol Mesa *et al.* 2003; **Tabla 2**) a este mismo alineamiento. El enfoque de MB es metodológicamente similar a Xdet, salvo que este último utiliza la clasificación funcional implícita en alineamiento en lugar de una clasificación externa. Como cabe esperar, los 6 residuos con los mejores valores de correlación predichos por MB no se agrupan alrededor del sitio activo (**Fig. 22b**). Sólo uno de ellos, el D97, coincide con los predichos por Xdet y MCdet. Se muestra así la importancia de utilizar métodos capaces de explotar información funcional conocida *a priori* (métodos supervisados) para los casos donde, incluso cuando esta clasificación corresponda con la implícita en el alineamiento, se sospeche que los detalles de la filogenia observada (es decir, sus distancias específicas) no son realistas.

III.III.3.D. Lactato / malato deshidrogenasas

Esta familia de enzimas homólogas comprende dos subfamilias de deshidrogenasas (ECs 1.1.1.27 y 1.1.1.37) que actúan sobre lactato y malato, respectivamente. En este caso, las proteínas están estrechamente relacionadas, sus relaciones en secuencia son claras y pueden ser alineadas por cualquiera de los métodos al uso.

La **figura 23a** muestra el árbol filogenético obtenido a partir del alineamiento múltiple de secuencias de la familia de lactato / malato deshidrogenasas (LDH y MDH, respectivamente) descrito en **Métodos Sección VI.VI.4**. Se observa que hay un grupo de MDHs (llamadas aquí MDH'), que está más cerca de las LDHs que del resto de MDHs. Un método no supervisado sería, en cierta medida, confundido por esta desavenencia, tratando de localizar

aquellos residuos comunes a LDHs y MDH', y diferentes del resto de MDHs, lo cual no reflejaría las diferencias de especificidad malato *versus* lactato.

En este caso, las clases funcionales corresponden a las actividades enzimáticas que difieren en su especificidad de sustrato (definida aquí como la diferencia en su último número del código EC, compartiendo los 3 primeros dígitos). En el método Xdet se utilizaron semejanzas funcionales binarias ("1" para los pares de proteínas que pertenecen a la misma clase y "0" para el resto).

La **figura 23b** muestra las predicciones de ambos métodos mapeadas sobre la estructura de la malato deshidrogenasa de *Aquaspirillum arcticum* (PDB 1b8u). En el caso de MCdet, se muestra la intersección entre las posiciones predichas para la especificidad malato y aquellas de la lactato. Estas deben corresponder con posiciones que tiendan a estar conservadas dentro de las MDHs y dentro de los LDHs pero con distinto tipo de aminoácido entre estas dos clases funcionales. Se observa que la mayoría de las posiciones, predichas de forma independiente por los dos métodos, están en el entorno del centro activo de la proteína (señalado en esta estructura por la unión con NAD y oxalacetato), si bien en este caso con menor cercanía que en los casos anteriores.

Esta familia de proteínas fue también estudiada en Hannenhalli y Russell (2000) para validar el método supervisado desarrollado en ese trabajo. Entre las predicciones arrojadas por Xdet y MCdet se encuentran dos de las seis posiciones detectadas por Hannenhalli y Russell (95 y 99 en la estructura PDB 1b8u, correspondientes a las 102 y 107 de su artículo). Las otras cuatro posiciones predichas por Hannenhalli y Russell mapean fuera de los límites del dominio Pfam sobre el que se basa el alineamiento utilizado, el cual cubre únicamente el dominio N-terminal de esta familia. Tanto MCdet como Xdet detectan la posición 95 (2 ª puntuación más alta de Xdet). Esta posición es una Arginina (R) en las MDHs y una Glutamina (Q) en las LDHs, habiendo una numerosas evidencias experimentales que demuestran su importancia en la determinación de la especificidad por lactato o malato (Hannenhalli y Russell, 2000). En este ejemplo, la relación entre las posiciones determinantes de especificidad y su cercanía al sustrato es menos evidente, e incluso la posición 95, determinada experimentalmente, no se encuentra en contacto con el sustrato (**Fig. 23b**).

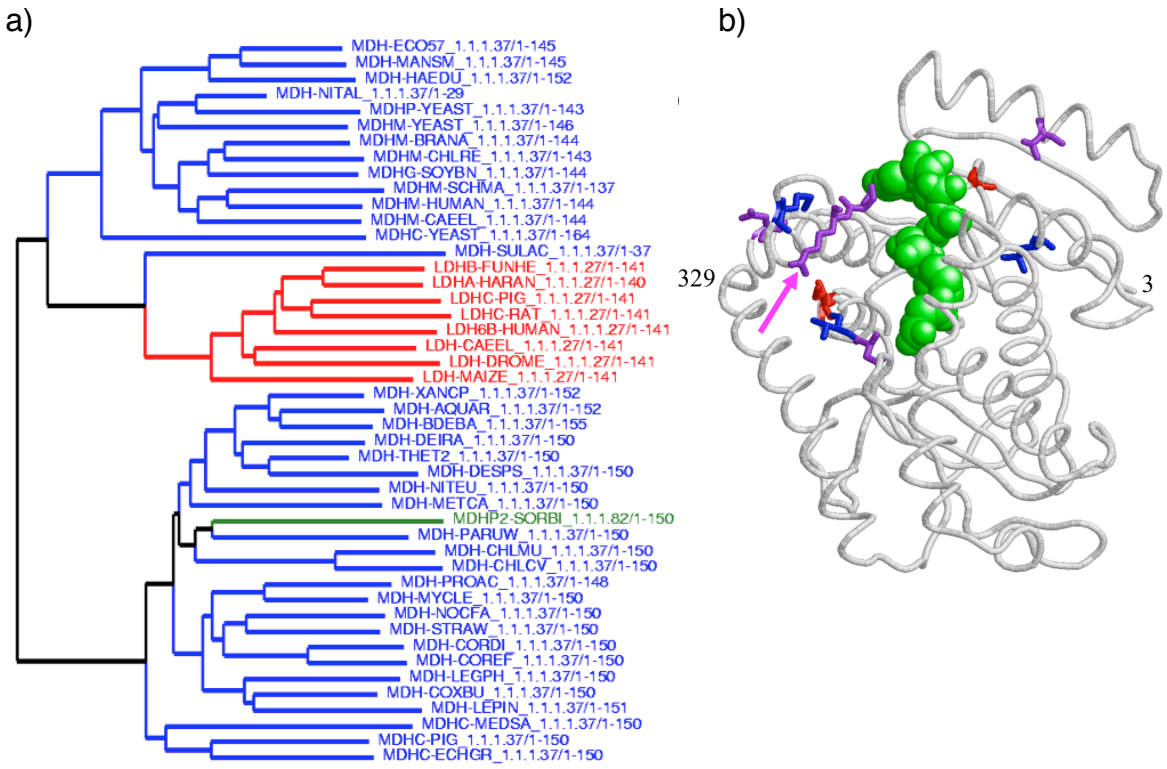


Fig. 23. Resultados de los métodos Xdet y MCdet para la familia de deshidrogenasas lactato / malato. (a) Árbol obtenido por unión de vecinos (*neighbour joining*) a partir del alineamiento descrito en Métodos Sección VI.VI.4. Las ramas del árbol y sus correspondientes proteínas están coloreadas de acuerdo a las diferentes clases funcionales a las que pertenecen: MDHs en azul y LDHs en rojo. La proteína coloreada en verde corresponde a una MDH que utiliza como cofactor NADPH en lugar de NADH. **(b)** Residuos predichos por los métodos Xdet y MCdet representados sobre la estructura malato deshidrogenasa de *A. arcticum* (PDB 1b8u). En la estructura, la proteína se encuentra unida a NAD y oxalacetato, representados en color verde. Los residuos predichos por los métodos se representan mediante bastones (*sticks*) de color azul para los del método Xdet, rojo para los del método MCdet y violeta para las posiciones predichas por ambos métodos. La flecha rosa indica la posición 95 (R en MDHs y Q en LDHs).

III.III.4. Implementación de MCdet en un software C/C++ distribuíble.

El protocolo descrito en MCdet se ha implementado en un software escrito en lenguaje C/C++. El programa es de libre uso para fines académicos. El paquete de distribución incluye el código fuente acompañado de archivos de ayuda para su instalación, configuración de variables locales, compilación de ejecutables, opciones de uso y manuales para la interpretación de la salida. Así mismo se incluyen diferentes ejemplos de inputs y sus correspondientes outputs que permiten la comprobación del correcto funcionamiento del programa una vez instalado. El programa se instala con éxito en diferentes arquitecturas bajo los sistemas operativos Linux y Mac OS X.

El método Xdet ha sido igualmente implementado en forma de software por el Dr. Florencio Pazos, el cual distribuye los ejecutables para uso académico bajo petición (ver la dirección web http://pdg.cnb.csic.es/pazos/Xdet/).

IV. Discusión

IV. DISCUSIÓN

En *El origen de las especies*, Darwin escribió: “Si se probara la existencia de un órgano complejo que no ha podido ser formado a partir de numerosas, sucesivas y ligeras modificaciones, mi teoría se vendría totalmente abajo”. La gran idea detrás de esta frase es que incluso las características biológicas “nuevas” portan el sello de sus ancestros. En el caso de las proteínas, la observación de características compartidas en secuencia y estructura permite inferir relaciones de homología, esto es, de descendencia a partir de una ancestro común (Fitch 1970). Las divergencia en secuencia entre proteínas homólogas ocurre por la acumulación de eventos de mutación, duplicación, especiación y deleción y se traduce en características funcionales distintivas adquiridas a escala evolutiva (Dayhoff *et al.* 1983). Sin embargo, esa divergencia no se observa en todas las posibles direcciones sino que está favorecida por la selección natural de estados intermedios funcionales que han moldeado el actual “espacio de proteínas” de los seres vivos (Maynard Smith 1970) con una estructura observable en superfamilias, familias y subfamilias de proteínas (Koonin *et al.* 2002; Yona *et al.* 1999).

Entre los eventos evolutivos anteriores, la duplicación génica se ha propuesto como el principal mecanismo que favorece la evolución de características funcionales entre proteínas homólogas, permitiendo explorar cambios funcionales por mutación en una de las copias mientras se mantiene la funcionalidad de la otra (Ohno 1999). En la ausencia de otros factores, se ha sugerido que las duplicaciones génicas tienen efectos neutrales (Lynch y Conery, 2003) o perjudiciales en el organismo (Wagner 2005; Gerstein *et al.* 2006; Bergthorsson *et al.* 2007), por lo que se esperaría que se perdieran por deriva genética o selección purificadora (Li 1980; Nei y Roychoudhury 1973; Force et al. 1999). Sin embargo, la amplia extensión de proteínas parálogas (homólogos resultantes de duplicación) observada incluso en los genomas de los organismos más sencillos, sugieren importantes ventajas selectivas conferidas por duplicación génica que han favorecido su fijación y conservación en las poblaciones respectivas (Conant y Wolf 2008). Estas ventajas incluyen la adquisición de características funcionales específicas, para las que se han propuesto dos modelos evolutivos principales: la subfuncionalización –el reparto de funciones ancestrales entre genes duplicados– y la neofuncionalización –adquisición de funciones en una de las copias y mantenimiento de la función ancestral en la otra– (Ohno 1970**;** Lynch 2000; Force et al. 1999; He y Zhang, 2005; Gibson y Goldberg, 2009).

En esta tesis investigamos la influencia de importantes aspectos de la especificidad funcional en la determinación de la divergencia en secuencia acumulada por una familia de proteínas homólogas. Bajo el marco conceptual que acabamos de exponer, la hipótesis de trabajo generalmente aceptada es que la divergencia en secuencia entre las subfamilias de proteínas está gobernada por una divergencia funcional (Devos y Valencia 2000; Rost 2002; Sjölander 2004). Asociadas a la organización interna de la familia, el alineamiento múltiple de sus proteínas (MSA)

permite observar ciertas posiciones con un patrón de conservación diferencial entre subfamilias (SDPs). Estos patrones reflejan presiones selectivas distintivas ejercidas sobre posiciones equivalentes entre las proteínas de las diferentes subfamilias y -extendiendo la hipótesis de trabajo mencionada- serían los responsables a nivel molecular de sus diferentes especificidades funcionales (ZuckerKandl y Pauling 1965).

El estudio computacional de la evolución de la especificidad funcional en una familia de proteínas a través la comparación de sus secuencias en un alineamiento múltiple conlleva una serie de limitaciones que pueden llegar a ser importantes en determinados casos donde el muestreo actual del espacio real de secuencias de la familia no sea representativo de su tamaño (grado de divergencia acumulado en la familia) , densidad (número de secuencias en el espacio) y estructura (organización interna en subfamilias). Los estudios basados en MSAs presentan también dificultades a la hora de considerar importantes aspectos estructurales (como las regiones desordenadas), evolutivos (como las inserciones y deleciones) o regulatorios (como las modificaciones post-traduccionales de sus residuos). Las regiones desordenadas y, en particular, los lazos (*loops*) de proteínas son funcionalmente importantes por su participación frecuente en el reconocimiento de ligandos y parejas de interacción (Xie *et al.*, 2007; Akiva et al., 2008; Stein *et al.*, 2009). Este tipo de regiones, por estar menos constreñidas estructuralmente, tiende a acumular una mayor de variabilidad (Brown *et al.*, 2002; Tompa 2003). Por otro lado, las inserciones y deleciones que sufren las proteínas en su evolución, participan también frecuentemente en las interfaces proteína-proteína, mediando o impidiendo interacciones específicas (Hashimoto y Panchenko, 2010; Akiva et al., 2008). Tanto las regiones desordenadas como las inserciones y deleciones son difíciles de integrar en los alineamientos múltiples de proteínas, tendiendo a introducir huecos (*gaps*) en el alineamiento. En consecuencia, este tipo de regiones -particularmente las desordenadas y las inserciones- tienden a ser descartadas y por tanto obviado su potencial rol funcional. También importante en la determinación de la especificidad funcional de una proteína es el conjunto de modificaciones post-traduccionales que pueden sufrir sus residuos (Seo y Lee, 2004). Estudios recientes señalan además el importante impacto que la adquisición diferencial de estas modificaciones puede tener en la evolución de las familias de proteínas (Amoutzias *et al.*, 2010; Stein *et al.*, 2009).

A pesar de las limitaciones anteriores, numerosos estudios computacionales basados en el estudio de MSAs han sido capaces de mostrar la relación entre la estructura de las subfamilias, SDPs e importantes aspectos de la especificidad funcional (**Introducción Sección I.11**) que incluyen principalmente la actividad catalítica propia de enzimas, la unión a pequeños ligandos y cofactores, y las interacciones con otras macromoléculas incluyendo otras proteínas, la unión a ADN/ARN, y otras. A mayor escala, considerando amplios conjuntos de familias, los estudios computacionales se han centrado en familias de enzimas y proteínas que unen pequeños ligandos (**Tabla 3**). Sin embargo, el estudio de la relación de subfamilias y SDPs con interacciones específicas proteína-proteína carecía de un

análisis sistemático en un conjunto de familias suficientemente representativo que permitiera dar generalidad a la asociación observada en casos particulares. En esta tesis se reevalúan los desarrollos anteriores integrándolos con el estudio de la relación entre subfamilias y SDPs con interacciones específicas proteína-proteína. Importantes aspectos de la función celular de las proteínas -como el control del ciclo celular, la señalización y el transporte- están mediados por interacciones entre proteínas y es de esperar que estas interacciones hayan condicionado su divergencia en secuencia en un grado importante. Por otra parte, conviene estudiar su influencia de forma integrada con la actividad catalítica y la unión a pequeños ligandos, considerando la capacidad de las interacciones proteína-proteína para alterar sus propiedades, crear nuevos sitios de unión, tener un papel regulatorio, etc.

La carencia anterior de un estudio como este podría explicarse en parte por la preponderancia de los estudios detallados a nivel bioquímico en proteínas individuales (enzimas principalmente) por contraste con el reciente auge de las biología de sistemas y, en consecuencia, el creciente interés en las interacciones proteína-proteína y sus redes derivadas. La expansión en el número de secuencias, estructuras, interacciones proteína-proteína y funciones anotadas nos ha permitido evaluar relaciones y conceptos sobre los que se ha estado discutiendo durante años en ausencia de datos brutos representativos de estos espacios. En esta tesis se presenta de forma integrada un análisis a gran escala de la organización interna de las familias de proteínas en términos funcionales y de la relación a nivel estructural que sus residuos característicos (SDPs) presentan con regiones funcionales fundamentales: sitios de unión a ligando e interfaces proteína-proteína. De especial interés es el estudio de la relación entre interacción diferencial entre proteínas y subfamilias y SDPs, el cual se presenta por vez primera de forma exhaustiva.

Se discuten a continuación los resultados obtenidos en cada una de las secciones específicas de este trabajo para pasar posteriormente a una discusión de sus implicaciones y las perspectivas que abren.

IV.I. El método S3det

Para la realización de este estudio se requería la utilización de una metodología escalable de análisis de MSAs en la que subfamilias y SDPs estuvieran definidas de forma coherente atendiendo a su mutua dependencia. Con ese objetivo se desarrolló una nueva metodología de análisis multivariante -el método S3det- inspirado en el abordaje de *Sequence Space* (Casari et al. 1995) que se presenta en la primera parte de esta tesis.

S3det permite definir subfamilias y SDPs de forma simultánea bajo un mismo marco metodológico basado en Análisis de Correspondencias Múltiple (MCA). La potencia de un abordaje multivariante como MCA reside en su capacidad para

desentramar las fuentes informativas de variabilidad de tal modo que la estructura en subfamilias y sus posiciones características surjan de forma simultánea. La lógica detrás de este abordaje explota la mutua dependencia de ambas entidades ya que las posiciones del MSA determinan la separación de las proteínas y al mismo tiempo esta separación pondera la contribución de las posiciones a esa segregación. Es esta una ventaja crucial respecto al resto de abordajes los cuales asumen, por lo demás, independencia matemática en la variación de las posiciones del MSA, asunción poco realista desde el punto de vista evolutivo (Tablas 2 y 3). Así pues, una metodología multivariante cobra especial sentido a nivel conceptual al tiempo que proporciona las herramientas matemáticas para llevar a cabo una detección concomitante de familias y SDPs asociados a estas.

Como se comentó en **Resultados Sección III.I.1**, el método *S3det* es conceptualmente similar al método *Sequence Space* (Casari *et al.* 1995) si bien hace uso de herramientas matemáticas esencialmente diferentes: *Sequence Space* está basado en el Análisis de Componentes Principales (PCA) mientras S3det utiliza el Análisis de Correspondencias Múltiple. La diferencia básica entre PCA y MCA es que el primero utiliza distancias euclídeas mientras que el segundo utiliza distancias chi-cuadrado. El uso de distancias chi-cuadrado permite la aparición de las relaciones pseudovaricéntricas entre los espacios de proteínas y residuos (**Resultados Sección III.I.1.E**) esto es: las coordenadas del centro de masas de las proteínas de una subfamilia coinciden exactamente con las coordenadas en el espacio de residuos de una posición del MSA que tenga un tipo de aminoácido exclusivamente conservado en esa subfamilia. Esto permite "mapear las subfamilias" en el espacio de residuos a través de su centro de masas y calcular distancias entre ellos. De este modo, la correspondencia entre los espacios de proteínas y residuos puede establecerse de modo analítico y por tanto ser automatizada sin ambigüedad. En PCA, por el contrario, la asociación matemática entre estos espacios presenta una serie de inconvenientes que han dificultado la automatización de este método y su aplicación a grandes conjuntos de proteínas. Así, en del Sol et al. (2003; método *SS-method*; **Tabla 2 y 3**) se logró automatizar la selección de SDPs en el espacio de residuos pero quedó por resolver la definición de forma automática de las subfamilias asociadas.

Como se ilustró en el ejemplo de la familia de aminotransferasas de clase III (**Fig. 4**), S3det es capaz de detectar patrones de conservación diferencial que impliquen bien a todas las subfamilias (esto es, conservación de un tipo de aminoácido diferente en cada una de las subfamilia), bien a cualesquiera partición disjunta y completa construida por agregación de las subfamilias detectadas (p.ej. conservación de un mismo tipo de aminoácido en dos subfamilias diferentes y conservación de un aminoácido distinto en el resto de subfamilias). El rastreo en todas sus posibles formas de la agregación de subfamilias está motivado por los siguientes dos aspectos:

En primer lugar, la función de una proteína está determinada por un conjunto de residuos que actúan de forma coordinada (Horovitz 1996). Así, puede ocurrir que dos subfamilias compartan un mismo tipo de aminoácido en una posición cuyo rol funcional (p.ej., en el caso de una enzima, la polarización de un enlace en un sustrato que va a ser roto) sea modulado por el resto de residuos que tienen en exclusiva (p.ej. reorientando su cadena lateral para determinar qué enlace del sustrato polariza dando lugar a productos diferentes). S3det toma en consideración estos aspectos y permite que existan -junto con SDPs cuyos aminoácidos son diferentes entre todas las subfamilias- otros que implican agregaciones de subfamilias. De esta manera se da especial valor a la noción de *conjunto* de posiciones determinantes de especificidad funcional, esto es: la especificidad funcional de una subfamilia dada aparece de forma natural explicada por una serie de residuos tomados como un todo, los cuales presentarán por un lado características exclusivas (aminoácidos exclusivos en esas posiciones respecto del resto de subfamilias) y por otro características distintivas aunque compartidas con otras subfamilias (aminoácidos comunes a varias subfamilias aunque diferentes de otras) (**Fig. 6**).

En segundo lugar, la agregación de subfamilias en todas sus posibles formas permite la detección de patrones de conservación diferencial en los cuales, para una determinada posición, se dé identidad de aminoácidos entre dos subfamilias cuya distancia en un árbol filogenético sea mayor que respecto a una tercera subfamilia con la cual difieran en el tipo de aminoácido. Esta posibilidad es especialmente interesante y diferencia a S3det respecto a los abordajes basados en árboles filogenéticos (p.ej. Evolutionary Trace) o en correlaciones con las distancias de las secuencias (p.ej. el método MB; **Tabla 2 y 3**) en los cuales se tiende a penalizan estos patrones.

El desarrollo de S3det nos ha permitido disponer de una metodología aplicable a un gran conjunto de familias a fin de obtener un número representativo de subfamilias y SDPs sobre el que evaluar, a continuación, el grado de su asociación con características funcionales específicas. Los resultados obtenidos mediante S3det son cualitativamente similares a los obtenidos mediante otros métodos para la predicción de SDPs (**Anexo I Sección A.I**): *Evolutionary Trace* (ET; Mihalek *et al.* 2004), *Combinatorial Entropy Optimization* (CEO; Reva *et al*., 2007), *Mutational Behavior* (MB; del Sol Mesa *et al*., 2003) y *Sequence Space* (Casari *et al*., 1995). Estos cuatro métodos son representativos del estado actual del campo y cubren los principales abordajes metodológicos a este respecto (**Tabla 2 y 3**). Si bien sus capacidades predictivas son similares, se pueden observar diferencias menores entre el comportamiento de los métodos. Así, en el ajuste de subfamilias a etiquetas funcionales, CEO tiende a ser más específico que S3det y Sequence Space (ET y MB no predicen subfamilias), aunque a costa de ser menos sensible. En terminos de la asociación con regiones estructurales, las predicciones de ET destacan sobre el resto, debido a que sus predicciones incluyen posiciones altamente conservadas (alrededor del 75% de sus predicciones están conservadas a más del 90% de identidad de secuencia en contraste con el ~1% del resto de métodos). Cuando se filtra la

conservación en niveles progresivos, los resultados de ET se asimilan paulatinamente a las otras metodologías.

La calidad de sus resultados unida a su capacidad para clasificar simultáneamente subfamilias y SDPs hacen de S3det un abordaje particularmente apropiado para los análisis propuestos en la segunda parte de esta tesis y que se discuten a continuación

IV.II. Estudio a gran escala de la contribución de importantes aspectos de la especificidad funcional a la evolución en secuencia de las familias de proteínas

Para el estudio de la contribución de importantes aspectos de la especificidad funcional a la evolución en secuencia de las familias de proteínas eucariotas se escogió como conjunto de análisis la base de datos Pfam (Bateman *et al*., 2004) de familias de dominios de proteínas. Pfam cubre un amplio espectro funcional, incluyendo dominios de función desconocida. Permite además trabajar con dominios en lugar de cadenas completas, lo cual es especialmente apropiado para este análisis por su carácter de unidad evolutiva básica (Apic *et al*., 2001; Chothia et al., 2003; Lee et al., 2003; Vogel et al., 2004; Kawashima et al., 2009). Sin embargo, trabajar con dominios de forma aislada presenta también inconvenientes como la posibilidad de estudiar sitios activos formados por la participación de dominios diferentes o por complejos de subunidades del mismo dominio dispuestas de forma asimétrica. Por otra parte, anotaciones funcionales como el código EC en el caso de los enzimas o la interacción con proteínas están establecidas sobre cadenas completas. Pese a interesantes estrategias para trasladar anotaciones funcionales a dominios (véase p.ej. Shoemaker y Panchenko 2007; López y Pazos 2009) esta limitación es difícilmente soslayable sin un estudio detallado de los casos.

Partiendo de esta base de datos, se aplicó S3det a un conjunto de 1262 familias de dominios de proteínas eucariotas obteniendo para ellas un amplio número de subfamilias y SDPs. Este conjunto ofrece la oportunidad de estudiar sistemáticamente y de forma integrada la distribución de secuencias y residuos clave en relación con dos aspectos cruciales de la función biológica de las proteínas: i) la función bioquímica correspondiente a su actividad catalítica y/o de unión a pequeños ligandos y ii) su unión específica a otras proteínas. Así, en el caso de las subfamilias de proteínas se analiza su correspondencia respecto a diferentes clases enzimáticas representadas por los códigos EC y diferentes conjuntos de interactores. En el caso de las SDPs, se estudia su distribución estructural respecto a sitios de unión a ligando, sitios catalíticos y regiones de interacción obtenidas a partir de información estructural.

Se establece de este modo un marco de análisis conceptualmente coherente de tal modo que los aspectos funcionales a estudiar pueden ser evaluados de forma

particular para cada una de las entidades objeto de estudio, esto es: subfamilias y SDPs. Al mismo tiempo, se permite someter a prueba la congruencia de las asociaciones funcionales de forma global, considerando por una parte la íntima relación entre las características funcionales utilizadas y, por otra, la mutua dependencia en la definición de subfamilias y SDPs. Los resultados de este estudio se discuten en las siguientes apartados específicos.

IV.II.1. Correspondencia entre subfamilias y etiquetas funcionales

Los resultados del análisis de la coincidencia entre subfamilias y etiquetas funcionales muestran una concordancia general entre la organización interna de las familias de proteínas y sus conjuntos de interactores específicos. Esta correspondencia se observa igualmente en el caso de la función bioquímica, si bien con valores de sensibilidad/especificidad acordes a los niveles esperables considerando la mayor caracterización de los códigos enzimáticos EC hasta la fecha (Devos y Valencia, 2000; Tian y Skolnick 2003) por contraste con la aún escasa caracterización de las interacciones proteína-proteína. El acuerdo entre subfamilias y códigos EC diferenciales es consistente con estudios previos también a gran escala (Wicker *et al.* 2001; Brown *et al.* 2007 ; Lee *et al*. 2009), aunque realizados mediante metodologías especializadas en la detección de subfamilias que no detectan SDPs. La capacidad de las subfamilias para reflejar conjuntos de interactores específicos, se muestra en cambio por primera vez de forma sistemática. De forma importante, los resultados obtenidos mediante las subfamilias obtenidas por métodos alternativos (**Anexo I Sección A.I**) soportan y dan generalidad al acuerdo global de las subfamilias de proteínas con ambos aspectos funcionales estudiados.

Como se comentó en Resultados (**Sección III.II.1**), la interpretación de los resultados obtenidos para ambas clasificaciones funcionales debe hacerse teniendo en cuenta sus respectivas naturalezas y tamaños, esto es: los grupos de proteínas para las cuales se ha determinado experimentalmente que interaccionan con la misma pareja (interactor) son de menor tamaño que aquellos grupos de proteínas etiquetados con el mismo código EC, los cuales tienden a ser -por contraste- grupos grandes. Cabe esperar también cierto sesgo debido a la asignación de códigos ECs por similitud en secuencia, difícilmente controlable por la ausencia en las bases de datos de registros claros en cuanto a la fuente (experimental o por homología) de la anotación.

No obstante, varias consideraciones pueden hacerse alrededor de los valores observados de sensibilidad/especificidad de las subfamilias respecto a los códigos EC y los conjuntos de interactores:

Por una parte, la clasificación enzimática EC puede englobar en su nivel descriptivo más detallado (el cuarto dígito) un rango más o menos amplio –según el caso- de especificidades (McDonald *et al.*, 2009; véase también Triviño y Pazos

2010). Por ejemplo la actividad enzimática “alcohol deshidrogenasa”, con código EC 1.1.1.1., engloba un espectro muy amplio de sustratos (alcoholes de muchos tipos). Así, la tendencia observada en los resultados de la **Fig. 12** (arriba) por la cual subfamilias diferentes pueden compartir el mismo EC, podría reflejar la abundancia relativa de estos casos cuyos detalles moleculares a un mayor nivel de especificidad estarían siendo recogidos por las diferentes subfamilias.

Por otra parte, en el caso de las interacciones diferenciales proteína-proteína, la clasificación en subfamilias obtenida parece recoger estos patrones de especialización funcional a un nivel de detalle más grueso que en el caso de la clasificación enzimática. Como se vio (**Fig 12**, abajo) las subfamilias detectadas tienden a contener proteínas con diferentes especificidades de interacción. Este resultado puede apuntar a la importancia en el establecimiento de interacciones específicas de otros niveles de regulación como el control de la expresión génica, los controles post-transcripcionales o la determinación de la localización subcelular. El papel de estos niveles de regulación en la determinación de los interactores específicos podría ser especialmente significativo cuando dichos interactores son parálogos entre sí, siendo probable que conserven su modo de interacción (Aloy *et al.* 2003) con diferentes proteínas de la misma subfamilia.

Por último, como se ha comentado anteriormente, cabe considerar que las etiquetas funcionales utilizadas para este estudio están establecidas sobre proteínas enteras, esto es, no desglosadas entre los dominios que las constituyen. Así, podría suceder que la especificidad funcional de una enzima o una interacción proteína-proteína no resida en el dominio que aquí se considere. Esto haría que la organización en subfamilias observada no refleje completamente las restricciones evolutivas de la función estudiada, lo cual podría explicar los valores mediocres de sensibilidad y especificidad obtenidos para algunas familias (**Fig. 12**).

Como valoración general de este apartado, la relación evidenciada entre subfamilias y especificidad funcional -tanto a nivel de interactores como a nivel bioquímico- da soporte cuantitativo a la hipótesis de divergencia en secuencia gobernada por divergencia funcional (Devos y Valencia, 2000; Rost, 2002; Sjölander 2004). Considerados de forma conjunta, los resultados del ajuste de subfamilias con etiquetas funcionales pueden interpretarse como una consecuencia de la naturaleza jerárquica de la diversificación funcional de acuerdo a los dos mecanismos principales propuestos para la evolución funcional de proteínas: la subfuncionalización (reparto de funciones ancestrales entre genes duplicados) y la neofuncionalización (mantenimiento de la función ancestral en una de las copias y adquisición de funciones en las otra) (He y Zhang, 2005; Gibson y Goldberg, 2009).

IV.II.2. Asociación estructural entre SDPs y regiones funcionales: sitios de unión a ligando e interfaces

Una vez establecida la relación de subfamilias con etiquetas funcionales, se caracterizó estructuralmente el conjunto de residuos (SDPs) asociados de forma robusta a la organización en subfamilias correspondiente. Para ello se analizó su distribución espacial respecto a un conjunto representativo de residuos de unión a pequeños ligandos y sitios catalíticos (López *et al.*, 2007) y de unión a proteínas identificados de forma fiable a partir de estructuras cristalográficas del *Protein Data Bank* (Berman *et al.*, 2000). Este tipo de regiones corresponden conceptualmente a las etiquetas funcionales analizadas en el caso de la distribución de secuencias en subfamilias: códigos EC y conjuntos de interactores específicos, respectivamente. Se completa de esta manera el diseño experimental esquematizado en la **Fig. 10**, considerando por una parte la íntima relación entre las características funcionales utilizadas y, por otra, la mutua dependencia en la definición de subfamilias y SDPs.

Los resultados obtenidos muestran una clara asociación de las SDPs con ambos sitios de unión, tanto en términos de distancias como en enriquecimientos estrictos. Como se comentó al inicio de esta discusión, la acumulación observada de SDPs en regiones de interacción proteína-proteína (**Fig. 15 y Fig. 17**) es especialmente novedosa y apunta al posible rol de los SDPs en la determinación de los modos de unión entre proteínas así como en la selección de parejas de interacción diferenciales entre subfamilias.

De forma complementaria, se estudió la distribución de las posiciones conservadas en cada familia respecto a las mismas regiones estructurales, mostrando que su asociación espacial es, en promedio, incluso más estrecha que la de las SDPs. Los resultados obtenidos para las posiciones conservadas son consistentes con estudios previos realizados también a gran escala que muestran un grado mayor de conservación en las interfaces de interacción que el resto de la superficie de la proteína (Choi *et al.* 2009; Engelen *et al.* 2009; Guharoy y Chakrabarti 2010).

La implicación de SDPs en interfaces de interacción extiende el uso de la conservación en el análisis de las interfaces a la consideración de las posiciones también conservadas pero de forma diferencial entre subfamilias. Así, las SDPs complementan a las posiciones mayoritariamente conservadas sobre las que hasta ahora se han analizado las interfaces y contribuyen a hacer más evidentes la señales de restricción evolutiva en secuencia en estas regiones funcionales.

IV.II.3. La asociación funcional de SDPs en proteínas que poseen tanto sitios de unión a ligando como regiones de interacción con proteínas.

Para entender mejor la relación entre sitios de unión a ligando y sitios de unión proteína-proteína, se ahondó en el estudio de las SDPs correspondientes a aquellas familias en las que se ha detectado experimentalmente ambos tipos de región funcional. Como resultado, se encontró un relevante número de familias en las que sus SDPs mapean en ambos tipos de regiones. Este resultado muestra la importancia de considerar de forma integrada diferentes aspectos funcionales. Así, los estudios de SDPs centrados exclusivamente en su relación con el sitio activo (**Tabla 3**) podrían haber subestimando la importancia funcional de aquellas SDPs implicadas en interacciones entre proteínas.

La presencia de SDPs en sitios activos e interfaces podría indicar bien una acción concertada de estos residuos en ambas funciones, bien una huella de eventuales mutaciones compensatorias que habrían permitido la evolución de una de las regiones funcionales de manera concertada con la otra. Esta situación ha sido ya resaltada anteriormente en el caso de la familia de aminotransferasas de clase III, donde las SDPs unen mediante contactos directos el sitio de unión a ligando con la interfaz de homodimerización. Cabe también señalar que el patrón de cambio coordinado dependiente de subfamilia de las SDPs podría solapar parcialmente con mutaciones correlacionadas intrafamilia, las cuales se han encontrado entre posiciones distantes y funcionalmente importantes dentro de una misma proteína (Fares y Travers 2006).

El estudio de las familias en las que se han detectado experimentalmente ambos tipos de región funcional muestra también un número importante de familias en las que alguno de sus SDPs foma parte simultáneamente de la interfaz y del sitio de unión a ligando. Así, de los 110 Pfams para los cuales existe solape entre ambos tipos de regiones funcionales, 33 familias (~30%; **Tabla 6**) presentan al menos uno de sus SDPs en la región de solape, las cuales pueden ser consideradas como posiciones “bifuncionales”. Recientemente, Davis y Sali (2010) han estudiado un conjunto de familias de estructuras de proteínas en las cuales se observa un solape significativo entre sus regiones de unión a ligando y las de unión a proteína. Los autores han caracterizado el nivel de conservación de estas posiciones bifuncionales (residuos que intervienen tanto en la unión al ligando como en la interfaz), mostrando que estas están menos conservadas que las “monofuncionales” (aquellas que intervienen en sólo un tipo de región funcional). Este resultado es coherente con los obtenidos en esta tesis (**Fig. 15**), en los que de forma complementaria se muestra la presencia de SDPs en estas regiones de solape y sugiere la posibilidad de que parte de las posiciones bifuncionales del estudio de Davis y Sali (2010) sean de hecho SDPs. Las posiciones “bifuncionales”, a cuya identificación podrían contribuir las SDPs en ausencia de información estructural, son de especial interés para el desarrollo de pequeñas moléculas con capacidad de modular interacciones proteína-proteína con

fines biotecnológicos o farmacéuticos, de forma complementaria a estudios computacionales -además del de Davis y Sali (2010)- que tratan de caracterizar los inhibidores de interacciones conocidos (Higueruelo *et al.* 2009) o predecir pequeñas moléculas inhibidoras a partir de interacciones con pequeños péptidos (Parthasarathi *et al.* 2008).

IV.II.4. Desglose de la asociación estructural de SDPs entre hetero-, homo- e intra-interfaces

A continuación se desglosó la distribución de SDPs de acuerdo a tres tipos de interfaces proteína-proteína: hetero, homo e intra-dominio. En los dos últimos casos, no se observó un enriquecimiento significativo de SDPs, si bien en un número importante de familias (**Fig 17**) sus SDPs se encuentran claramente implicados en regiones de homodimerización y, en casos específicos como las aminotransferasas de clase III anteriormente discutidas, aparecen potencialmente implicados en la modulación de la especificidad funcional. Las posiciones conservadas (definidas al 90% de identidad de secuencia) sí aparecen en cambio enriquecidas de forma significativa en interfaces de los 3 tipos, tanto homodiméricas como intradominio, de forma consistente con los resultados obtenidos en estudios anteriores (Jones y Thornton, 1996; Jones y Thornton, 1997; Choi *et al.* 2009; Engelen *et al.* 2009; Guharoy y Chakrabarti 2010). Por el contrario, se observa una implicación específica de SDPs en interfaces heterodiméricas, a un nivel similar -también significativo, aunque menor que en los casos anteriores- al observado para las posiciones conservadas en este mismo tipo de interfaces. Desafortunadamente, el número de familias de proteínas con miembros de sus diferentes subfamilias cristalizados en homocomplejos y heterocomplejos no es aún suficiente para llevar a cabo un estudio más detallado y con el suficiente respaldo estadístico. De forma importante, los resultados obtenidos mediante las SDPs definidas por métodos alternativos (**Anexo I Sección A.I**) soportan de forma robusta la asociación de SDPs con interfaces heterodiméricas, si bien la caso de las interfaces homo- e intra- esta asociación sigue sin ser concluyente.

Junto al enriquecimiento estadístico, los resultados obtenidos muestran que las interfaces proteína-proteína presentan un amplio rango de variabilidad en su enriquecimiento en SDPs. La heterogeneidad observada muestra la importancia de disponer de un conjunto suficientemente amplio y representativo para el estudio que aquí se presenta, a la vez que alerta sobre las diferencias que podrían arrojar estudios con conjuntos de prueba menores.

Como valoración general de este apartado, la consideración en conjunto de los resultados estadísticos obtenidos unida a la observación de los casos individuales, permite proponer como uno de los mecanismos de evolución de la unión específica a proteínas la selección de residuos clave cuya conservación diferencial entre

subfamilias determinaría la correspondiente unión específica a sus efectores (véase p.ej. el caso de la familia de factores de transcripción E2F/TDP expuesto en **Resultados Sección III.I.2.B**).

IV.III. Desarrollo de dos metodologías para la predicción de residuos funcionales utilizando información funcional supervisada

El estudio a gran escala discutido en la sección anterior da soporte cuantitativo a la hipótesis por la cual la divergencia en secuencia entre las subfamilias de proteínas está gobernada por una divergencia funcional. Bajo este escenario general, las relaciones filogenéticas de las proteínas de una familia (representadas por sus parecidos relativos en secuencia) se traducen en una organización interna en subfamilias ligada a determinadas posiciones con un patrón de conservación diferencial que reflejan características funcionales específicas. En consecuencia, la detección de los residuos determinantes de la especificidad funcional en una familia de proteínas debería poder realizarse en la mayoría de casos mediante métodos no supervisados (**Tabla 2 y 3**) basados en el análisis de la divergencia en secuencia observada a través de su alineamiento múltiple.

Como se vio en la Introducción **(Sección I.10)**, se han desarrollado también diferentes metodologías para la predicción de SDPs de forma supervisada (**Tabla 4**), esto es: cuando se dispone *a priori* de una clasificación en subfamilias. Estos métodos han demostrado también su capacidad de detectar posiciones funcionalmente importantes (véase Capra et al. 2008 para una evaluación comparativa), si bien su rango de aplicación biológica (como sugiere esta tesis) no es esencialmente diferente del de los métodos no supervisados. De hecho, la motivación principal para el desarrollo de métodos supervisados ha sido la necesidad de soslayar las limitaciones de los no supervisados para establecer subfamilias en determinados casos (**Tabla 2**).

No obstante, pueden encontrarse casos específicos donde los métodos supervisados sean de utilidad por haber un desacuerdo entre la clasificación basada en secuencia y la clasificación funcional *de facto* (p.ej. en algunos alineamientos estructurales donde se comparan homólogos muy lejanos o en casos excepcionales de convergencia evolutiva). También trabajos recientes como los de Halabi *et al.* (2009) y Schwarz *et al.* (2009), encuentran casos en los que diferentes “sectores” del alineamiento muestran una divergencia específica correspondiente a diferentes restricciones funcionales. En esos casos, las proteínas de la familia pueden agruparse de diferentes formas según el aspecto funcional estudiado. En cambio, si la agrupación en subfamilias se realiza sobre la secuencia completa, esas agrupaciones alternativas quedarían enmascaradas por la filogenia “compuesta”.

Esta tesis se complementa con dos metodologías supervisadas (MCdet y Xdet) que se desarrollaron cuando el repertorio de métodos supervisados de referencia se

limitaba a los de Hannenhalli y Russell (2000) y Mirny y Gelfand (2002). La novedad de MCdet en ese momento estribaba en su capacidad para predecir posiciones con un tipo de aminoácido conservado para un determinado grupo preestablecido de proteínas pero variable en el resto. Este tipo de posiciones pueden reflejar un cambio brusco en la restricción funcional ejercida sobre esa posición (Gu 2001). Las posiciones conservadas exclusivamente en un grupo pueden en cambio no ser importantes si la divergencia entre las proteínas de ese grupo es reciente, pero la posibilidad de que reflejen una restricción evolutiva con implicaciones funcionales aumenta cuando se dan entre proteínas con homología remota que han de ser comparadas mediante alineamientos estructurales (algunos de los cuales son precisamente el terreno de juego de los métodos supervisados tal como ilustran los ejemplos estudiados en **Resultados Sección III.III.3**). Posteriormente se han desarrollado otros métodos capaces de predecir este tipo de posiciones (Chakrabarti *et al.* 2007; Capra y Singh 2008) si bien, a diferencia de MCdet, se han evaluado en escenarios típicos de métodos no supervisados. Así pues, la caracterización de la capacidad predictiva de residuos funcionales de MCdet sigue siendo de interés. Por su parte Xdet continua siendo desde su publicación el único método supervisado diseñado para explotar información funcional cuantitativa (p.ej. constantes enzimáticas, de afinidad de unión, etc).

La ventaja de *Xdet* con respecto al resto de métodos, incluido *MCdet*, es su capacidad de explotar información cuantitativa relativa a semejanzas o jerarquías funcionales (en contraste con el resto de métodos, los cuales requieren clases funcionales disjuntas). Se trata de una característica importante considerando la amplitud del concepto "función proteica" que requiere complejas clasificaciones jerárquicas y ontologías para ser codificado (Harris *et al.* 2004). Como principal limitación, *Xdet* es capaz únicamente de predecir posiciones con una importancia global en la determinación del conjunto de especificidades (esto es, no está diseñado para predecir posiciones responsables exclusivamente de una especificidad determinada). Como contrapartida, al buscar posiciones con una importancia global, *Xdet* puede trabajar con un número relativamente pequeño de proteínas de cada subclase.

El método MCdet es una simplicación del método S3det presentado en la primera parte de esta tesis (si bien fue desarrollado con anterioridad). A diferencia de S3det, MCdet trabaja exclusivamente en el espacio de residuos y sobre él proyecta un vector función que codifica de forma binaria la pertenencia a una función determinada de las proteínas del alineamiento. MCdet presenta las ventajas de los abordajes multivariantes, esto es: MCdet permite establecer distancias entre el vector función estudiado y los patrones de ausencia/presencia de un tipo de aminoácido en una posición de las secuencias del MSA, teniendo en cuenta la mutua dependencia entre secuencias y posiciones en el alineamiento. Como desventaja potencial, *MCdet* hace un tratamiento cualitativo de los tipos de aminoácidos a diferencia de otros métodos -

como Hannenhalli y Russell (2000), Mirny y Gelfand (2002) y el propio Xdet- que incorporan información sobre la similitud entre los aminoácidos.

Como se ha dicho, la capacidad predictiva de residuos funcionales de MCdet y Xdet se evaluó en diferentes alineamientos de proteínas para los cuales se comprobó cuidadosamente que la clasificación funcional disponible no estuviera implícita en la filogenia. Para ello se muestran cuatro ejemplos que abarcan un amplio rango de relaciones entre secuencias (desde la proximidad considerable hasta la basada puramente en alineamientos estructurales) en los que se explotan diferentes definiciones de función y con diferentes formas de cuantificar los parecidos funcionales. En estos ejemplos los métodos no supervisados no sería de aplicación, como se ilustra explícitamente en el caso de las hidrolasas glicosídicas de estructura barril TIM (**Resultados Sección III.III.3.C**). En esta familia las distancias en secuencia entre las proteínas son muy elevadas, haciéndolas poco fiables para la predicción de residuos funcionales mediante un método que explota sus parecidos relativos. Así, las predicciones obtenidas mediante un método no supervisado (MB, del Sol *et al.* 2003 **Tabla 2**) no tienen una relación espacial con las regiones funcionales, a diferencia de las obtenidas mediante Xdet y MCdet que se agrupan alrededor del centro activo.

IV.IV. Implicaciones y perspectivas

El conjunto de resultados presentados en esta tesis amplía el marco funcional en el que se han estudiado las subfamilias y SDPs para considerar -como aspectos igualmente relevantes en su evolución- las interacciones proteína-proteína de forma integrada con la actividad catalítica y la unión a pequeños ligandos. La asociación mostrada entre subfamilias y conjuntos diferenciales de interactores unida a la distribución de SDPs en interfaces proteína-proteína, tiene una serie de implicaciones, tanto prácticas como conceptuales. Discutimos en esta sección estas implicaciones y las perspectivas de trabajo que abren.

En primer lugar, la presencia observada de SDPs tanto en sitios de unión a ligando como en interfaces sugiere la posibilidad de que exista una comunicación alostérica entre estas posiciones para la determinación de la especificidad funcional. Existe evidencia de que residuos lejanos a la interfaz pueden jugar un rol determinante en la estabilización del complejo (Hedstrom 1996). También proteínas como las de la superfamilia Ras, sufren un cambio conformacional desencadenado desde el sitio activo (de unión a GTP en este caso) que las habilita para el reconocimiento de efectores específicos (Colicelli 2004, Wennerberg et al. 2005). Se ha propuesto que este tipo de residuos estarían conectados energéticamente con aquellos situados directamente en la interfaz, permitiendo que la energía de unión se propague a lo largo de la estructura terciaria de la proteína (Lockless y Ranganathan, 1999). Trabajos como los de Süel *et al*., (2003), del Sol *et al*. (2006) y Baussand y Carbone (2009) sugieren que redes de residuos conectados físicamente y que están

evolutivamente conservados podrían constituir motivos estructurales cuya función sería la comunicación alostérica en el seno de la proteína. Se abre pues la atractiva posibilidad de que las SDPs, además de su participación tanto en el sitio de unión a ligando como en la interfaz, estén implicadas en la propagación de señales alostéricas determinantes del reconocimiento y unión de interactores.

Por otra parte, como se ha comentado, la definción de SDPs y subfamilias es mutuamente dependiente. En consecuencia, la implicación de SDPs en interfaces es coherente y se complementa con la capacidad de las subfamilias para reflejar conjuntos de interactores específicos observada anteriormente. Esta coherencia puede ponerse en relación con la capacidad de otros abordajes computacionales en los que el análisis de la coevolución en secuencia entre familias permite detectar interacciones físicas proteína-proteína (Pazos y Valencia 2001; Pazos *et al.* 2005; Juan *et al.* 2008) así como predecir pares de interacciones específicas entre las proteínas de dos familias (Ramani y Markote 2003; Izarzugaza *et al.* 2006). En ese contexto, las SDPs de dos familias cuyas proteínas interaccionan reflejarían posiciones con un patron de cambio coordinado en la interfaz de unión que, a su vez, determinaría la interacción específica y diferencial entre las proteínas de las diferentes subfamilias. De hecho, esta posibilidad ha sido ya investigada por Chackrabarti y Panchencko (2009) sobre un pequeño conjunto de pares de proteínas interactoras. En este trabajo los autores señalan la importancia de las SDPs como posiciones responsables de la coevolución entre familias en la que sus proteínas interaccionan de forma específica.

En el mismo sentido que en el párrafo anterior, la presencia de SDPs en la interfaz unida a la posibilidad de que las SDPs de las proteínas interactoras presenten un patron de cambio coordinado, se asemeja a la distribución de “mutaciones correlacionadas” en las regiones de interacción proteína-proteína puesta de manifiesto por diferentes trabajos (Pazos *et al.* 1997; Hernanz-Falcón *et al.* 2004 ;Tress *et al.* 2005; Yeang y Haussler 2007; Madaoui y Guerois 2008). La señal de este tipo de posiciones podría en estos casos solapar parcialmente con la de las SDPs, cuando los cambios correlacionados correspondan a la organización en subfamilias.

Así mismo, la asociación de SDPs con interfaces heterodiméricas abre la posibilidad de la utilizar SDPs para guiar soluciones de ensamblaje computacional (*docking*; Tress e*t al.*, 2005). La utilización en las dinámicas moleculares de residuos funcionales predichos a partir de información de secuencia junto con información estructural previa es prometedora a juzgar por los últimos trabajos de los grupos de Szurmant Hoch y Terence Hwa en la Jolla (Schug *et al.,* 2009; Weigt e*t al.*, 2009). En estos trabajos se utilizan las posiciones con cambios correlacionados entre las familias de las proteínas que se quiere ensamblar para guiar las soluciones de las dinámicas moleculares. Las SDPs podrían contribuir en este tipo de estrategia aportando de forma adicional información sobre la localización de la interfaz. Esta posibilidad está siendo ya explorada en nuestro grupo en colaboración con los

doctores Ludovicco Sutto y Francesco Gervasio del grupo de Biofísica Computacional del CNIO.

Por otra parte, en esta tesis se ha visto que la asociación de SDPs con interfaces de heterocomplejos no sólo ocurre en la interfaz "final" sino que se observa también en términos de distribución de distancias a la interfaz. Estos resultados pueden ponerse en relación con el trabajo reciente de Wass *et al.* 2011, donde los autores muestran que la distribución de las puntuaciones obtenidas mediante un modelo de acoplamiento entre proteínas (*docking*) es capaz de distinguir interactores conocidos respecto de un fondo de proteínas que no interaccionan. En el mismo trabajo se muestra también la posibilidad de identificar interacciones partiendo de un conjunto de proteínas más parecidas pertenecientes a la misma superfamilia. Los autores sitúan estos resultados en el contexto del modelo de embudo del paisaje de energías intermoleculares en el ensamblaje de las interacciones proteína-proteína (McCammon 1998; Tsai *et al.* 1999). Según esta teoría, la interacción entre dos proteínas comienza con encuentros inespecíficos seguidos de reorganizaciones en su orientación favorecidas progresivamente por contactos más fuertes y específicos. En consecuencia, no sólo la interfaz final contiene información para la interacción sino también otras partes de la superficie de la proteína. Bajo este modelo, las SDPs podrían estar implicadas en los diferentes modos de unión que transcurren hasta la interfaz definitiva y que –en los resultados de Wass et al.- explicarían que la distribución de energías moleculares de las interfaces intermedias sea informativa. Esta posibilidad está siendo ya investigada en nuestro grupo.

En relación con el párrafo anterior, la implicación de las SDPs en la especificidad de las interacciones proteína-proteína heterodiméricas es también compatible con el modelo propuesto por el cual se sitúa la colocalización de proteínas (y el subsiguiente incremento de su concentración local) en el origen y evolución diferencial de las interacciones proteicas (Kuriyan y Eisenberg 2007). La colocalización de dos proteínas puede surgir por mecanismos diversos entre los que se encuentran la compartimentación subcelular, la asociación a membranas, la unión a ADN o ARN y la fusión génica (pasando a formar parte de la misma cadena polipeptídica). El modelo de Kuriyan y Eisenberg propone que la colocalización de dos proteínas incrementa la concentración local de una respecto a la otra aumentando la selectividad de aquellas mutaciones que afectan a la estabilidad del complejo. Las SDPs aparecen como una consecuencia natural en el contexto de este modelo, pudiendo corresponder a los residuos bajo una presión selectiva diferencial reflejo de la diversidad de situaciones en las que esa colocalización haya tenido lugar. Así, sería interesante explorar cómo la asociación de subfamilias y SDPs con interacciones específicas se corresponde y modula con su colocalización en diferentes orgánulos celulares así como con la correlación de sus niveles de expresión en función del tipo celular, tejido o estadío de desarrollo.

En esta tesis se ha visto también que, si bien la asociación de SDPs en interfaces homodiméricas no es suficientemente significativa para un amplio conjunto de familias, su implicación en este tipo de regiones es observable en un número considerable de casos (**Fig. 17**). En el caso de este tipo de interfaces, los estudios a gran escala de homooligómeros han mostrado que la mayoría de familias con este tipo de interacción exhiben un único modo de unión conservado a lo largo de todas sus proteínas cristalizadas (Levy et al., 2008; Dayhoff *et al.*, 2010). Sin embargo, existe también un número de familias en las cuales el modo de homooligomerización varía entre sus miembros. Cuando la simetría varía a lo largo de las proteínas de una misma familia, Levy y colaboradores (2008) han mostrado rutas bien definidas por las cuales se evoluciona de estructuras cuaternarias más simples a otras más complejas. Esta evolución ocurre de forma progresiva y jerárquica: la más reciente sirve de base para la aparición de la siguiente, con prevalencia –cuando existe la opción– de las innovaciones diédricas (aparición de contactos simétricos "cara-a-cara" o "espalda-a-espalda") sobre las cíclicas (aparición de contactos "cara-a-espalda"). Por su parte, como se comentó en la introducción (**Sección I.11**) Dayhoff *et al.* (2010) han estudiado la evolución de nueve familias en las cuales el modo de homo-oligomerización varía entre sus miembros, relacionando sus diferentes modos de unión en los respectivos árboles filogenéticos. De forma interesante, estos autores han observado que los modos de unión entre homo-oligómeros (y sus correspondientes simetrías) tienden a estar conservados dentro de las diferentes subfamilias y de forma diferencial entre ellas. La combinación de los estudios de Levy et al., 2008 y Dayhoff *et al.*, 2010 sugiere la hipótesis de que las posiciones diferencialmente conservadas entre subfamilias (SDPs) sean las responsables de sus diferentes modos de homomerización. Así, las SDPs podrían representar las innovaciones progresivas ocurridas en la evolución de la familia que determinarían los contactos cara-a-cara o cara-a-espalda responsables de las nuevas formas de simetría.

Otra importante cuestión que se abre a partir del trabajo presentado en esta tesis es la implicación evolutiva de las SDPs en aspectos de la especificidad funcional determinados por una contribución epistásica entre residuos funcionales, esto es: residuos cuyos efectos fenotípicos no se explican de forma aditiva sino sinérgica. Este escenario ha sido ilustrado recientemente por Gloor et al. (2010) a partir de las características de crecimiento de cepas de levadura con diferentes pares de mutaciones realizadas sobre la fosfoglicerato-quinasa endógena PGK1. Los autores muestran situaciones en las que las características de crecimiento de los dobles mutantes no se explican por los efectos aditivos de los mutantes simples. En cambio, las posiciones estudiadas muestran comportamientos epistásicos entre los residuos implicados. Si se observa ahora el esquema representado en la **Fig. 24** se verá cómo el patrón de conservación de SDPs que implican diferentes agregaciones de subfamilia (véase p.ej. las posiciones 86 y 99 del esquema) ofrece una situación análoga a la de mutantes simples y dobles. Siguiendo esta analogía, las SDPs de nuestro estudio podrían reflejar variantes "naturales" simples y dobles. La contribución

aditiva o sinérgica de estos pares de posiciones a la especificidad funcional de las proteínas implicadas podría ser evaluada en aquellos supuestos para los que se disponga de información funcional de tipo cuantitativo (p.ej. constantes enzimáticas o de afinidad de unión de pequeñas moléculas o proteína-proteína).

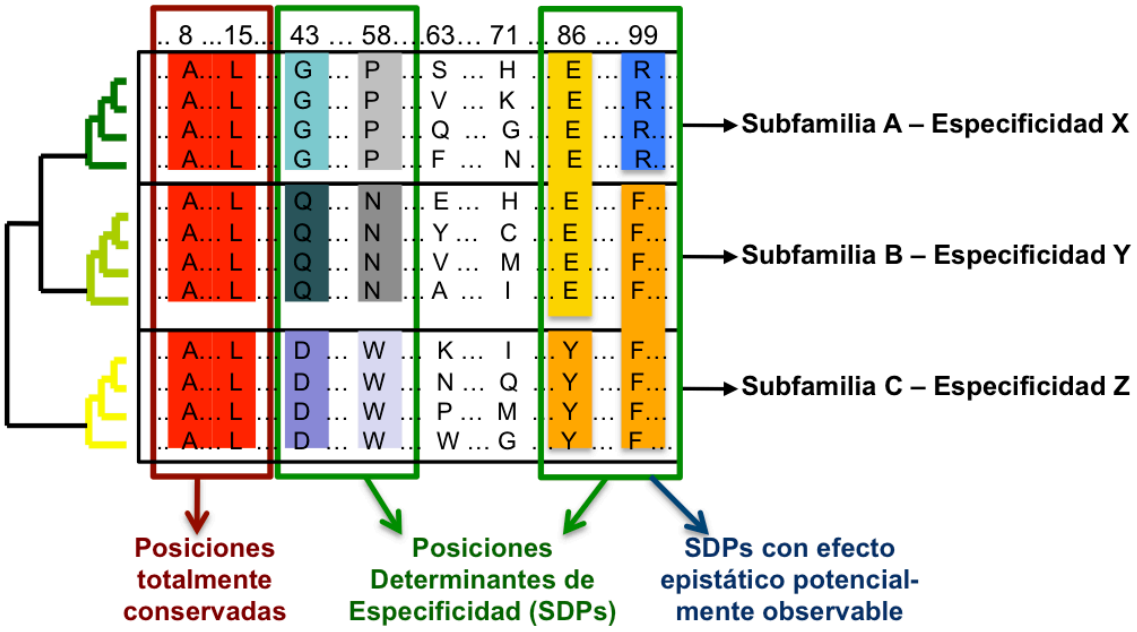


Fig. 24. Representación esquemática de un hipotético alineamiento múltiple de proteínas. El esquema ilustra cómo el patrón de conservación de SDPs que implican diferentes agregaciones de subfamilia (véase p.ej. las posiciones 86 y 99 del esquema) ofrece una situación análoga a la de mutantes simples y dobles

Además de en el estudio de la especificidad funcional desde una perspectiva evolutiva, los fenómenos epistásicos entre residuos de una proteína son importantes para comprender la variabilidad genómica en el seno de una población en relación a sus diferencias fenotípicas. Así, numerosos estudios de re-secuenciación están tratando de identificar polimorfismos entre humanos que puedan asociarse al desarrollo de enfermedades como el cáncer (GWAs, del inglés *Genome Wide Association Studies*) (Goldstein DB 2009). Si bien esta estrategia ha dado resultados positivos, el éxito ha sido menor del que se esperaba (Galvan et al. 2010), por lo que diferentes autores señalan la importancia de estudiar la asociación de polimorfismos con enfermedades teniendo en cuenta los posibles efectos epistásicos que puedan tener (Daly y Altshuler, 2005), esto es: no de forma aditiva sino sinérgica considerados en conjunto. En el caso particular de polimorfismos en un solo núcleotido (SNPs del ingés *Single Nucleotide Polymorphisms*) que ocurren en regiones codificantes de una proteína, la estrategia metodológica seguida en este trabajo para la identificación de SDPs puede aplicarse para identificar conjuntos de SNPs asociados de forma

conjunta al desarrollo de una enfermedad. Esta estrategia es análoga a la representada en la **Fig. 24,** en la que en lugar de proteínas homólogas de diferentes organismos se alinean las secuencias de una misma proteína en diferentes individuos. De forma interesante, ambas estrategias pueden ser combinadas de tal modo que se prioricen aquellas asociaciones de SNPs en posiciones que son además SDPs en la familia correspondiente. De hecho, la utilizad de las SDPs para predecir mutaciones causantes de enfermedad está ya siendo explotada con éxito (Chris Sander; comunicación personal).

Como valoración final, esta tesis contribuye a la comprensión de la contribución de importantes aspectos funcionales a la evolución en secuencia de las familias de proteínas. El estudio integrado de las interacciones de proteínas junto con la actividad catalítica y la unión a pequeñas moléculas extiende el marco conceptual de subfamilias y SDP hacia una definición más comprehensiva de función que, a su vez, abre nuevas hipótesis en torno a otros importantes aspectos evolutivos relacionados con la especificidad funcional en proteínas.

V. Conclusiones

V. CONCLUSIONES

1. La organización en subfamilias de las familias de proteínas responde de forma general a características funcionales diferenciales relacionadas con la actividad enzimática específica y con conjuntos distintivos de proteínas interactoras.

2. Las posiciones diferencialmente conservadas en subfamilias (SDPs) están estructuralmente asociadas a regiones funcionales correspondientes a sitios catalíticos, sitios de unión a ligando e interfaces proteína-proteína, tanto en términos de distribuciones de distancias espaciales como en enriquecimientos relativos.

3. Las familias de proteínas para las cuales se dispone de información estructural tanto de sitios de unión a ligando como de interfaces proteína-proteína, presentan sus SDPs en uno, otro o ambos tipos de regiones funcionales de forma cualitativamente similar sin observarse una tendencia preferente por alguna de ellas.

4. Cuando el estudio del enriquecimiento en interfaces se desglosa en tres tipos correspondientes a interacciones heterodiméricas, homodiméricas e intracadena, se observa que las SDPs están significativamente enriquecidas en interfaces que implican heterocomplejos. Sin embargo, este enriquecimiento no se observa en un número significativo de familias para el resto de interfaces (homodiméricas e intracadena) cuando estas se consideran de forma aislada.

5. Las conclusiones anteriores se reproducen utilizando las subfamilias y/o SDPs definidas por metodologías alternativas. El método S3det desarrollado en esta tesis produce unos resultados cualitativamente similares al resto de metodologías ensayadas, con la ventaja de que la definición de subfamilias y SDPs se hace de forma simultánea y coherente.

6. Los resultados estadísticos obtenidos unidos a la observación de los casos individuales, permiten proponer como uno de los mecanismos de evolución de la unión específica a proteínas la selección de residuos clave cuya conservación diferencial entre subfamilias determinaría la correspondiente unión específica a sus interactores.

7. Los métodos Xdet y MCdet desarrollados son capaces de explotar información funcional cuantitativa (Xdet) y clasificaciones supervisadas (MCdet) para predecir residuos determinantes de especificidad funcional en alineamientos de proteínas donde los parecidos en secuencia de las proteínas no se corresponden con las similitudes funcionales.

8. Consideradas en conjunto, las conclusiones anteriores dan generalidad y soporte cuantitativo a la hipótesis por la cual la divergencia en secuencia acumulada en las familias de proteínas está gobernada por su divergencia funcional.

VI. Material y Métodos

VI. MATERIAL y MÉTODOS

VI.I. Obtención del conjunto de alineamientos de proteínas de la base de datos Pfam.

Se partió de la base de datos completa Pfam-A de alineamientos múltiples de secuencias de familias de proteínas (Bateman *et al.*, 2004, versión 22.0) sobre los que se realizaron los siguientes filtros en términos tanto de familias como de secuencias y posiciones (aunque sin realinear los MSAs originales):

1. Sólo se consideran secuencias de organismos eucariotas para las cuales existe evidencia a nivel de proteína o de tránscrito de acuerdo a la información ofrecida por la base de datos UniProtKB/Swiss-Prot (Apweiler *et al.* 2004, versión 54.6)

2. En aras a trabajar únicamente con el área compacta de los alineamientos, éstos se restringen a aquellas secuencias y a aquellas posiciones con menos de un 30% de huecos (*gaps*). Ésta selección se realiza de forma recursiva como se sintetiza a continuación en forma de pseudocódigo:

   para (i=70; i>=30; i=i-10) { mantener aquellas secuencias con (% de huecos <= i) calculado sobre aquellas posiciones con (% de huecos <= i) }

   Para cada 'i', se recuperan las secuencias Pfam-A originales de modo que, en este punto, su versión sin huecos permanece idéntica a la inicial.

3. Los alineamientos se hacen no redundantes al 95% en identidad de secuencia. La redundancia se eliminó utilizando en programa Cd-hit (Li y Godzik, 2006). Para cada grupo redundante de secuencias se mantiene una única secuencia representante entre aquellas que maximizan el siguiente esquema de puntuación: i) si tiene un código PDB asociado: +4; ii) si tiene código EC asociado: +2; si es una secuencia de humanos: +1. En caso de empate se elige la más larga.

4. Las secuencias atípicas (*outliers*) se eliminan del análisis. Se define aquí "secuencia atípica" como aquella secuencia con un porcentaje de identidad sobre la región alineada menor del 40% con respecto a cualquier otra del alineamiento.

5. Se eliminan aquellas posiciones con un porcentaje de huecos mayor del 10%

6. Únicamente se consideran alineamientos con al menos 12 secuencias y al menos 25 posiciones). Por otra parte, no se han impuesto límites superiores ni en el número de secuencias ni en el de posiciones en los alineamientos analizados.

VI.II. Construcción de clases funcionales y análisis de la organización funcional en subfamilias

V.II.1. A partir de información enzimática

La clasificación EC procede de la base de datos UniprotKB (Apweiler et *al.* 2004). La selección de familias de Pfam con suficiente información funcional relativa al código EC se realizó como sigue:

Se definen "clases de proteínas" dentro de una familia como aquellas que comparten los 2 primeros dígitos EC (p.ej. EC 2.11. / EC 2.10. / … se consideran, todas ellas, clases diferentes entre sí) y "grupos de proteínas" como los subconjuntos dentro de clase que comparten los 4 dígitos (p.ej. EC 2.11.7.5 / EC 2.11.7.8 / EC 2.11.3.1 se consideran todos ellos grupos diferentes dentro de la misma clase). Las familias de Pfam que presentan "clases" parcialmente solapantes o "grupos" dentro de clase parcialmente solapantes se descartan del análisis. Dada una familia Pfam con varias clases no solapantes, sólo se considera aquella clase que represente a un mayor número de secuencias y, en caso de igualdad, aquella que contenga una mayor diversidad de grupos. Únicamente se consideran aquellas familias que cumplen simultáneamente los siguientes requisitos: i) contienen más de un grupo de ECs con al menos 3 secuencias cada uno; ii) éstos grupos contienen al menos el 25% de las secuencias de la familia; iii) el grupo de mayor tamaño contiene menos del 80% de las secuencias que pertenecen a algún grupo.

VI.II.2. A partir de interacciones proteína-proteína

Los pares de proteínas interactoras en los organismos *S. cerevisiae* y *H. sapiens* se extrajeron de los experimentos etiquetados como "pequeña escala" en el núcleo (*core*) de la base de datos DIP (Xenarios *et al.* 2000, a fecha de 14-10-2008 para *S. cerevisiae* y 26-01-2009 para *H. sapiens*). Para cada uno de estos dos organismos se construyó un conjunto de negativos (pares de proteínas que no interaccionan) a partir de los pares de proteínas para las que se cumple que: i) ambas proteínas aparecen anotadas manualmente en la base de datos KEGG (Kanehisa *et al.* 2004; versión 49.0) perteneciendo a diferentes rutas metabólicas; o ii) no comparten localización celular, habiendo sido éstas determinadas experimentalmente de acuerdo a las bases de datos MIPS (Mewes *et al.* 1997; a fecha de 14-11-2005) para *S. cerevisiae*, y de la base de datos eSLDB (Pierleoni et *al.* 2007; a fecha de 21-1-2009) para ambos

organismos. Las localizaciones celulares consideradas -y su equivalencia correspondiente entre estas dos últimas bases de datos- son las siguientes:

Base de datos MIP	Base de datos eSLDB
701 extracellular	Extracellular
705 bud	(sin categoría
710 cell wall	Cell wall
715 cell periphery	(sin categoría
720 plasma membrane	(sin categoría
725 cytoplasm	Cytoplasm
730 cytoskeleton	Cytoskeleton
735 ER	Endoplasmic reticulum
740 golgi	Golgi
745 transport vesicles	Vesicles
750 nucleus	Nucleus
755 mitochondria	Mitochondrion
760 peroxisome	Peroxisome
765 endosome	Endosome
770 vacuole	Vacuole
775 microsomes	Lysosome

Tabla 8. Categorías de localización subcelular utilizadas y su equivalencia entre las bases de datos MIPS y eSLDB (ver texto)[1]**.**

Como requisito adicional en la definición de negativos, se exige siempre que el par de proteínas no se encuentre anotado como positivo en BIOGRID (Stark *et al.* 2006; versión 2.0.49), repositorio general de conjuntos de datos de proteínas interactoras que incluye experimentos masivos.

VI.II.3. A partir de información funcional implícita en los identificadores de proteínas SwissProt ID

Se considera que dos proteínas pertenecen al mismo grupo "ID" si la primera parte de su identificador coincide (p.ej. OAT_HUMAN, OAT_MOUSE, OAT_YEAST, etc.). La primera parte del identificador SwissProt ID representa

[1] Las categorías se citan en inglés a fin de evitar ambigüedades respecto a las empleadas en las bases de datos citadas.

una abreviación del nombre del gen/proteína correspondiente (representando el organismo al que corresponden mediante la segunda parte del identificador).

Se considera que una familia Pfam dispone de suficiente diversidad funcional en términos de identificadores SwissProt ID si cumple los siguientes requisitos: (i) posee más de un grupo de IDs con al menos 3 secuencias por grupo; (ii) al menos el 25% de las secuencias del alineamiento están incluídas en algún grupo de IDs; y (iii) ningún grupo de IDs representa más del 80% de las secuencias con grupo de ID asignado.

VI.III. Obtención de la información estructural sobre sitios de unión a ligando e interfaces proteína-proteína

VI.III.1. Obtención de conjuntos Pfam estructuralmente redundantes a nivel de superfamilia de SCOP

Cada secuencia de un MSA para la cual existe información estructural disponible se alinea de forma apareada contra sus correspondientes estructuras cristalográficas del PDB, obtenidas a través de la base de datos MSD (Velankar *et al.* 2005). Únicamente se consideran PDBs procedentes de experimentos de difracción de rayos X con una resolución menor o igual a 3 Ångströms y con un máximo de 10 átomos no determinados. Los alineamientos apareados se realizan mediante el programa Blas2Seq (Altschul *et al.* 1997).

Entre estas estructuras, se seleccionan aquellas con un dominio estructural asignado por SCOP (Andreeva *et al.* 2004) que solape al menos un 80% (de forma recíproca) en el alineamiento contra el correspondiente dominio Pfam. Por otra parte, aquellas familias Pfam cuyas secuencias mapean a más de una superfamilia de SCOP se descartan del análisis.

Finalmente, el conjunto de familias Pfam que mapean sobre la misma superfamilia de SCOP se considera un “conjunto estructuralmente redundante”. Los conjuntos así generados se emplean en los “análisis no redundantes a nivel estructural”, a fin de evitar sesgos debidos a una representación desigual de las superfamilias de SCOP en el conjunto de familias Pfam de partida.

VI.III.2. Obtención de los sitios de unión a ligando e interfaces proteína-proteína

En las estructuras seleccionadas en la sección anterior, los residuos funcionales se recopilaron como sigue:

Los sitios catalíticos y de unión a ligando se extrajeron de la base de datos FireDB (López *et al*. 2007) desarrollada en nuestro grupo. FireDB integra datos de los contactos atómicos derivados de estructuras PDB así como los

residuos catalíticos anotados de forma fiable en el *Catalytic Site Atlas* (Porter *et al.* 2004).

Para la definición estructural de sitios de unión proteína-proteína (interfaces) se utilizaron únicamente las unidades biológicas (archivos BioUnit PDB) disponibles en la base de datos RCSB PDB (a fecha de 03-12-2008; ftp://ftp.rcsb.org/pub/pdb/data/biounit/coordinates; Berman *et al.*, 2000). El conjunto de complejos de PDB se extraen de la base de datos 3D complex (a fecha de 25-5-2008; Levy *et al.* 2006). De éstos, se descartan aquellos complejos anotados como errores según la iniciativa de supervisión manual PiQSi (a fecha de 25-5-2008; Levy 2007). Las cadenas de proteínas con menos de 60 residuos se eliminaron de los correspondientes PDBs a fin de evitar interacciones con fragmentos de proteínas e interacciones péptido-proteína.

Los sitios de interacción proteína-proteína se definieron de acuerdo al criterio estándar basado en el cambio de accesibilidad tras la interacción (Valdar y Thornton 2001). Los residuos en superficie se definen como aquellos con una superficie relativa accesible (RSA*,* del inglés relative accessible surface area) igual o mayor al 5%. La RSA se calcula mediante el programa NACCES (Valdar y Thornton 2001). Los residuos en la interfaz se definen como aquellos que cumplen el criterio de accesibilidad cuando la cadena se considera de forma aislada pero no cuando el cálculo se realiza sobre el complejo, ésto es, que se hacen inaccesibles con la unión. Estos cálculos se realizan a través del programa pdb_defineface (http://www.biochem.ucl.ac.uk/bsm/valdarprograms/pdb_defineface.html) de los mismos autores, el cual utiliza a su vez el programa NACCES citado anteriormente. Siguiendo las recomendaciones, los archivos BioUnit PDB se filtraron previamente a través del programa pdb_fixcoord disponible en http://www.biochem.ucl.ac.uk/bsm/valdarprograms/pdb_fixcoord.html con los siguiente parámetros: *--rm_nocalpha --het_nonaminos –unhet_aminos*.

Las interfaces se clasifican en *homo*-, *hetero*- e *intra*-interacciones según impliquen dos cadenas representando la misma proteína (de acuerdo a su código Swissprot AC), dos proteínas diferentes, o dos dominios estructurales de la misma proteína (de acuerdo a la definición estructural de dominio provista por SCOP). Las interfaces homoméricas y heteroméricas se definieron a pares de cadenas. P.ej., para un archivo BioUnit PDB con tres cadenas A, B y C, la interfaz en A se calcula como la suma de la de A con B más la de A con C. Sólo se contemplan aquellos residuos de la secuencia Pfam que se encuentren entre los límites SCOP de dominio estructural a menos que la longitud del resto de la cadena fuera del correspondiente límite sea inferior a 15 residuos.

VI.III.3. Obtención de un PDB representante de cada familia Pfam y mapeo sobre él de los residuos funcionales

Para cada MSA se elige únicamente una de sus estructuras como representante óptimo de la familia Pfam y sobre ella se proyecta la información estructural del resto (residuos de unión a ligando y constitutivos de la interfaz). La selección del representante óptimo se realiza a través de la infraestructura implementada en FireDB. En FireDB, todas las cadenas de PDB se agrupan en conjuntos redundantes al 97% de identidad de secuencia, para cada uno de los cuales se genera su secuencia consenso llamada "secuencia máster". En primer lugar, para cada familia Pfam, se selecciona su "secuencia máster representante" como aquella que maximice, sucesivamente, los siguientes parámetros:

1. La suma total del "% de SDPs" más el "% de residuos de unión a ligando acumulado a lo largo de todos los PDBs presentes en la familia" que cubre.
2. La cobertura del alineamiento (sin considerar columnas con un porcentaje de huecos del 10%).
3. El porcentaje de identidad de secuencia con su secuencia de referencia.

Se elige entonces una sóla cadena PDB -de entre las que aglutina la secuencia máster elegida en el punto anterior- como aquella que maximice:

1. La suma total del "% de PDSs" más el "% del total residuos de unión a ligando acumulado a lo largo de todos los PDBs presentes en la familia" que cubre.
2. La resolución (en Ångströms) del experimento de difracción de rayos X del que procede.

VI.IV. Cálculo de distancias entre residuos de estructuras PDB

Las distancias entre residuos de estructuras cristalográficas PDB calculadas en este trabajo corresponden a distancias euclídeas medidas en Ångströms a partir de las coordenadas de los residuos disponibles en los archivos anteriormente citados. Las distancias se calculan entre las coordenadas de los carbonos β de cada residuo excepto cuando se trate de una Glicina, en cuyo caso se utiliza el carbono α.

VI.V. Tests de enriquecimiento

Los conjuntos de SDPs y posiciones conservadas (definidas al 90% de identidad de secuencia) se comparan en este trabajo con los conjuntos de residuos funcionales recopilados (de unión a ligando y de unión a proteínas) mediante los

llamados tests de enriquecimiento. Este tipo de test trata de evaluar la sobrerepresentación de residuos funcionales en un determinado conjunto de residuos y a lo largo de un número significativo de casos por encima de lo que es esperable por azar. Así, para cada familia estudiada se calcula el porcentaje de SDPs que están anotados como residuos funcionales (según la definición aquí empleada) y la proporción correspondiente de éstos últimos en toda la proteína. El enriquecimiento se define como la diferencia de éstos dos porcentajes para cada familia; sobre la lista total de éstos se realiza un test de Wilcoxon para datos apareados (Wilcoxon 1945). Éste test calcula el p-valor de la hipótesis nula que establece que no existe diferencia significativa entre la representación relativa de residuos funcionales en nuestro subconjunto frente a cualquier otro tomado al azar. El procedimiento es análogo cuando se evalúan el subconjunto de residuos conservados así como cuando se desglosa el tipo de residuos funcionales entre aquellos de unión a ligando o a proteína y éstos últimos entre *hetero*-, *homo*- e *intra*-interfaces.

VI.VI. Alineamientos de proteínas estudiados mediante los métodos supervisados Xdet y MCdet

VI.VI.1. Homólogos estructurales del oncogén Ras

Se partió del alineamiento estructural generado automáticamente por el programa Dali (Holm y Sander, 1994) a partir de la estructura tridimensional del oncogén Ras (PDB 1ctqA). El alineamiento obtenido contiene proteínas de unión a diferentes tipos de ligando, incluyendo nucleótidos (GTP, FMN, FAD ...), nucleósidos, azúcares, etc. El alineamiento se filtra dejando sólo las cadenas con una similitud estructural con la máster (1ctqA) superior a 6,0 (puntuación ZFSSP), se elimina redundancia por encima del 40% de identidad de secuencia, y se descartan las estructuras que no presentan unión a ligando. El alineamiento final contiene 24 proteínas unidas a diferentes tipos de ligandos.

Como medida de similitud funcional para el método Xdet se utiliza aquí la similitud química entre las moléculas unidas dada por el coeficiente de Tanimoto (Holliday *et al.* 2002). Éstos se obtuvieron para cada par de ligandos desde el servidor SuperLigands (http://bioinf.charite.de/superligands). Para los casos en los que alguna de las proteínas comparadas une a más de un ligando, se escogen los ligandos entre los cuales hay mayor semejanza. El solvente y las moléculas pequeñas no funcionales se excluyen del análisis.

Todos los nucleótidos de guanina (GTP, GDP y GNP) se consideran como una sola clase (referida como "GxP") dado que las propiedades de unión de las proteínas a las que unen son idénticas y el hecho de que se encuentre unidas a un nucleótido de guanina en particular se debe a factores extrínsecos a la

proteína (p.ej. su cristalización con nucleótidos artificiales no hidrolizables, etc.). Así, dos proteínas que unan GTP y GNP respectivamente, son consideradas aquí funcionalmente idénticas. El método *MCdet* se aplicó en este caso para la predicción de los residuos responsables de esta función "GxP" (unión de núcleotidos de guanina).

VI.VI.2. Dominios SH3

La similitud funcional entre las clases de dominios SH3 mencionadas (**Sección VI.VI.II**) con las que se ejecutó el programa *Xdet*, se derivó en este caso a partir de una clasificación funcional jerárquica de los dominios SH3 desarrollada por expertos (Cesareni et *al.* 2002). En el contexto del método *Xdet*, la distancia funcional entre dos clases dadas se define aquí, a partir de este árbol funcional, como el número de ramas entre esas dos clases y el primer nodo interno en el que convergen. La similitud funcional (*sim*) se obtiene de la distancia funcional (*dist*) como sim = max_dist - dist, donde max_dist es la máxima distancia (4 en este caso). Por ejemplo, *sim* (*1R*, *1R*) = 4 (la más alta en este árbol); *sim* (*1R*, *2R*) = 3, *sim* (*1R*, *2D*) = 0, etc. El método *MCdet* se aplicó en la predicción de los residuos asociados con la clase funcional *1R* (la única con suficientes representantes en el alineamiento).

VI.VI.3. Hidrolasas glicosídicas con estructura de barril TIM

Se partió del alineamiento estructural generado automáticamente por el servidor *Dali* a partir de la estructura PDB 1qumA. El alineamiento se filtra dejando sólo las cadenas con una similitud estructural con la máster (1qumA) superior o igual a 7,0 (puntuación ZFSSP) y se elimina redundancia por encima del 30% de identidad de secuencia. El alineamiento resultante está compuesto de estructuras de barril TIM (*TIM-barrels*), la mayoría de las cuales son enzimas pertenecientes a diferentes clases EC. Se restringió el alineamiento a las hidrolasas glicosídicas (o glicosidasas, de código EC 3.2.1.*.) El alineamiento final contiene con 20 secuencias pertenecientes a un total de 10 subclases diferentes de glicosidasas (3.2.1.1, 3.2.1.2, 3.2.1.35, ...).

En este caso, se utilizaron semejanzas funcionales binarias en el método *Xdet*, esto es: si las proteínas A y B pertenecen a las misma subclase de hidrolasas se les asigna una similitud igual a 1, y 0 en el caso contrario. El método *MCdet* fue utilizado para predecir las posiciones responsables de la subclase 3.2.1.1, la única con suficientes representantes en el alineamiento.

VI.VI.4. Lactato / malato deshidrogenasas

Se partió del alineamiento Pfam (Bateman *et al.* 2004) con código PF00056 del dominio de unión a NAD de las lactato/malato deshidrogenasas. Este dominio comprende los residuos 1 al 145 de la proteína malato deshidrogenasa de *Escherichia coli*. Se filtró la redundancia por encima del 80% de identidad de secuencia. El alineamiento final contiene un total de 46 proteínas.

VI.VII. Cálculo de árboles filogenéticos

Los árboles filogenéticos de los conjuntos de proteínas anteriores se obtuvieron, a partir de los alineamientos descritos para cada caso, mediante el programa *Belvu* (Sonnhammer y Hollich, 2005 ; disponible en http://sonnhammer.sbc.su.se/Belvu.html) con el método *neighbour-joining* (unión de vecinos) utilizando los parámetros por defecto.

VII. Bibliografía

VII. BIBLIOGRAFIA

Abascal F, Valencia A (2002) Clustering of proximal sequence space for the identification of protein families. *Bioinformatics* 18:908-921

Akiva E, Itzhaki Z, Margalit H (2008) Built-in loops allow versatility in domain–domain interactions: Lessons from self-interacting domains. *PNAS* 105(36):13292-13297

Aloy P, Ceulemans H, Stark A, Russell RB (2003) The Relationship Between Sequence and Interaction Divergence in Proteins. *J Mol Biol* 332(5):989-998

Aloy P, Querol E, Aviles FX, Sternberg MJE (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol* 311: 395-408

Aloy P, Russell RB (2002) Interrogating protein interaction networks through structural biology. *PNAS* 99(9):5896-5901

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402

Amoutzias GD, He Y, Gordon J, Mossialos D, Oliver SG, Van de Peer Y. Posttranslational regulation impacts the fate of duplicated genes. *PNAS* 107(7):2967-2971

Andrade MA, Casari G, Sander C, Valencia A (1997) Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol Cybern* 76:441-450

Andreeva A, et al. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 32:D226-229.

Apweiler R, Bairoch A, Wu CH, *et al.* (2004) UniProt: The universal protein knowledgebase. *Nucleic Acids Res* 32:D115–D119

Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser SN, Orchard S, Perreau V, Roechert B, van Eijk K, Hermjakob H. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res* 38:D525-D531

Armon A, Graur D, Ben-Tal N (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* 307: 447-463

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1):25-29

Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, Zygouri C (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 31(1):400-402.

Bairoch A. (1992) PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res* 20: 2013-2018

Bateman A, Coin L, Durbin R*, et al*. (2004) The Pfam protein families database. *Nucleic Acids Res* 32:D138-141

Bauer B, Mirey G, Vetter IR, Garcia-Ranea JA, Valencia A, Wittinghofer A, Camonis JH, Cool RH (1999) Effector recognition by the small GTP-binding proteins Ras and Ral. *Journal of Biological Chemistry* 274:17763-17770

Baussand J, Carbone A (2009) A Combinatorial Approach to Detect Coevolved Amino Acid Networks in Protein Families of Variable Divergence. *PLoS Comput Biol* 5(9)

Beltrao P, Serrano L (2007) Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput Biol* 3:e25

Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* 20(8):1322 -1324

Bergthorsson U, Andersson DI, Roth JR. (2007) Ohno's dilemma: evolution of new genes under continuous selection. *PNAS* 104:17004–17009

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235-242

Bickel PJ, Kechris KJ, Spector PC, Wedemayer GJ, Glazer AN (2002) Finding important sites in protein sequences. *PNAS* 99:14764-14771

Blake C. (1983) Exons and the evolution of proteins. *Trends Biochem. Sci.* 8:11-13.

Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. *J Mol Biol* 280:1-9

Bordner AJ, Abagyan R. (2005) Statistical analysis and prediction of protein-protein interfaces. *Proteins* 60(3):353-366

Brenner,S.E. A tour of structural genomics. Nat. Rev. Genet. (2001), 2: 801–809.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. Evolutionary rate heterogeneity in proteins with long disordered regions (2002) *J Mol Evol* 55(1):104-110

Brown DP, Krishnamurthy N, Sjölander K (2007) Automated Protein Subfamily Identification and Classification. *PLoS Comput Biol* 3(8):e160

Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci* 13(1):190-202

Calinski T, Harabasz J (1974) A Dendrite Method for Cluster Analysis. *Comm Stat* 3(1):1-27

Capra JA, Singh M (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics* 24(13):1473-1480

Capra JA, Singh M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics* 23(15):1875-1882

Carro A, Tress M, de Juan D, Pazos F, Lopez-Romero P, del Sol A, Valencia A, Rojas AM. (2006) TreeDet: a web server to explore sequence space. *Nucleic Acids Res* 34:W110-W115

Casari G, Sander C, Valencia A. (1995) A method to predict functional residues in proteins. *Nat Struct Biol* 2:171-178.

Cesareni G, Panni S, Nardelli G, Castagnoli L (2002) Can we infer peptide recognition specificity mediated by SH3 domains? *FEBS Lett* 513:38-44

Chakrabarti S, Bryant SH, Panchenko AR (2007) Functional Specificity Lies within the Properties and Evolutionary Changes of Amino Acids. *J Mol Biol* 373(3):801-810

Chakrabarti S, Panchenko AR (2009) Coevolution in defining the functional specificity. *Proteins* 75(1):231-240

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Research* 35(Database):D572-D574

Chelliah V, Chen L, Blundell TL, Lovell SC (2004) Distinguishing Structural and Functional Restraints in Evolution in Order to Identify Interaction Sites. *J Mol Biol* 342(5):1487-1504

Choi YS, Yang J-S, Choi Y, Ryu SH, Kim S (2009) Evolutionary conservation in multiple faces of protein interaction. Proteins 77(1):14-25

Chothia C, Lesk AM. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J* 5:823–826

Clackson T, Wells JA (1995) A hot-spot of binding-energy in a hormonereceptor interface. *Science* 267:383-386

Colicelli J. (2004) Human RAS Superfamily Proteins and Related GTPases. Sci STKE 2004(250):RE13-RE13

Conant GC, Wolfe KH. (2008) Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet* 9(12):938-950

Cusick ME, Yu H, Smolyar A, Venkatesan K, Carvunis A-R, Simonis N, Rual J-F, Borick H, Braun P, Dreze M, Vandenhaute J, Galli M, Yazaki J, Hill DE, Ecker JR, Roth FP, Vidal M (2009) Literature-curated protein interaction datasets. *Nat. Methods* 6(1):39-46

Daly MJ, Altshuler D (2005) Partners in crime. *Nat Genet* 37(4):337-338

Davis FP, Sali A (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* 21(9):1901-1907

Dayhoff JE, Shoemaker BA, Bryant SH, Panchenko AR (2010) Evolution of Protein Binding Modes in Homooligomers. *J Mol Biol* 395(4):860-870

Dayhoff MO, Barker WC, Hunt LT (1983) Establishing homologies in protein sequences. *Methods Enzymol* 91:524-45

De Hoon MJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. *Bioinformatics* 20(9):1453-1454

del Sol A, Fujihashi H, Amoros D, Nussinov R (2006) Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol Syst Biol* 2:2006.0019-2006.0019.

Del Sol Mesa A, Pazos F, Valencia A (2003) Automatic Methods for Predicting Functionally Important Residues. *J Mol Biol* 326(4):1289-1302

DeLano WL (2002) Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol* 12:14-20

Dessimoz C, Boeckmann B, Roth AC, Gonnet GH (2006) Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res* 34(11):3309-3316.

Devos D, Valencia A (2000) Practical limits of function prediction. *Proteins* 41:98-107

Donald JE, Shakhnovich EI (2005) Determining functional specificity from protein sequences. *Bioinformatics* 21(11):2629-2635

Eddy SR (1996) Hidden Markov Models. *Curr Opin Struct Biol* 6:361-365

Elcock AH (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J Mol Biol* 312(4):885-896

Engelen S, Trojan LA, Sacquin-Mora S, Lavery R, Carbone A (2009) Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput Biol* 5(1):e1000267.

Fares M, Travers SAA (2006) A novel method for detecting intramolecular coevolution: adding a further dimension to select constraints analyses. *Genetics* 173: 9–13

Fariselli P, Casadio R (2001) RCNPRED: prediction of the residue co-ordination numbers in proteins. *Bioinformatics* 17:202–204

Fariselli P, Olmea O, Valencia A, Casadio R (2001) Prediction of contact maps with neural networks and correlated mutations. *Protein Eng* 14:835-843

Felsenstein, J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164-166

Feng Z, Chen L, Maddula H, Akcan O, Oughtred R, Berman HM, Westbrook J (2004) Ligand depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* 20:2153–2155

Fields S, Song O. (1989) A novel genetic system to detect protein-protein interactions. *Nature* 340(6230):245-246

Finn RD, Marshall M, Bateman A (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 21(3):410-412

Fischer JD, Mayer CE, Söding J (2008) Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics* 24(5):613 -620

Fischer TB, Arunachalam KV, Bailey D, *et al.* (2003) The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics* 19:1453–1454.

Fitch WM (1970) Distinguishing homologous from analogous proteins. *Syst. Zool* 19(2):99-113

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269(5223):496-512

Fodor AA, Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 56(2):211-221

Force A *et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545

Frank R (1992) Spot-synthesis: An easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. *Tetrahedron* 48:9217–9232.

Fujimoto Z, Takase K, Doui N, Momma M, Matsumoto T, Mizuno H (1998) Crystal structure of a catalytic-site mutant alpha-amylase from Bacillus subtilis complexed with maltopentaose. *J Mol Biol* 277:393-407

Galvan A, Ioannidis JPA, Dragani TA. (2010) Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends Genet* 26(3):132-141

Gavin AC, Bösche M, Krause R, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141-147

Gerstein AC, Chun HJ, Grant A, Otto SP (2006) Genomic convergence toward diploidy in Saccharomyces cerevisiae. *PLoS Genet* 2, e145

Gibson TA, Goldberg DS (2009) Questioning the Ubiquity of Neofunctionalization. *PLoS Comput Biol* 5(1):e1000252

Gilbert W. Genes-in-pieces revisited (1985) *Science* 228(4701):823-824.

Gloor GB, Tyagi G, Abrassart DM, Kingston AJ, Fernandes AD, Dunn SD, Brandl CJ. (2010) Functionally Compensating Coevolving Positions Are Neither Homoplasic Nor Conserved in Clades. *Mol Biol Evol* 27(5):1181-1191

Göbel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18:309-317

Goldstein DB. (2009) Common genetic variation and human traits. *N. Engl. J. Med* 360(17):1696-1698

Golemis EA, Tew KD, Dadke D (2002) Protein interaction-targeted drug discovery: evaluating critical issues. *BioTechniques* 32(3):636-638, 640, 642 passim

Greenacre M, Blasius J (2006) Multiple correspondence analysis and related methods. *Springer Berlin Heidelberg*

Greenacre MJ (1984) Theory and Application of Correspondence Analysis. London Academic Press

Grishin NV, Phillips MA (1994) The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci* 3(12):2455-2458

Gu X (2001) Maximum-Likelihood Approach for Gene Family Evolution Under Functional Divergence. *Molecular Biology and Evolution* 18(4):453 -464

Guharoy M, Chakrabarti P (2010) Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *BMC Bioinformatics* 11(1):286

Guharoy M, Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. *PNAS* 102(43):15447-15452

Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes H-W, Stümpflen V (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res* 34:D436-441

Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res* 31(1):371-373

Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* 138(4):774-786

Halperin I, Wolfson H, Nussinov R (2004) Protein-Protein Interactions: Coupling of Structurally Conserved Residues and of Hot Spots across Interfaces. Implications for Docking. *Structure* 12(6):1027-1038

Halperin I, Wolfson H, Nussinov R. (2006) Correlated mutations: Advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins: Structure, Function, and Bioinformatics* 63(4):832-845

Hannenhalli SS, Russell RB (2000) Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol* 303:61-76

Harris MA, Clark J, Ireland A et al. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32:D258-D261

Hashimoto K, Panchenko AR (2010) Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *PNAS* 107(47):20352 - 20357

He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157-1164

Hedstrom L (1996) Trypsin: a case study in the structural determinants of enzyme specificity. *Biol. Chem* 377(7-8):465-470

Hendlich M, Bergner A, Gunther J, Klebe G (2003) Relibase: Design and developement of a database for comprehensive analysis of protein-ligands interactions. *J Mol Biol* 326:607–620

Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L (1997) Gene families: the taxonomy of protein paralogs and chimeras. *Science* 278(5338):609-614

Hernanz-Falcón P, Rodríguez-Frade JM, Serrano A, Juan D, del Sol A, Soriano SF, Roncal F, Gómez L, Valencia A, Martínez-A C, Mellado M (2004) Identification of amino acid residues crucial for chemokine receptor dimerization. *Nat. Immunol* 5(2):216-23

Higueruelo AP, Schreyer A, Bickerton GRJ, Pitt WR, Groom CR, Blundell TL (2009) Atomic Interactions and Profile of Small Molecules Disrupting Protein–Protein Interfaces: the TIMBAL Database. *Chemical Biology & Drug Design* 74(5):457-467

Ho Y, Gruhler A, Heilbut A, *et al.* (2004) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415(6868):180-183

Holliday JD, Hu CY, Willett, P. (2002) Grouping of coefficients for the calculation of intermolecular similarity and dissimilarity using 2D fragment bit-strings. *Comb Chem High Throughput Screen* 5:155-166

Holm L, Sander C (1994) The FSSP database of structurally aligned protein fold families. *Nucl. Acids Res* 22:3600-3609

Horovitz A (1996) Double-mutant cycles: A powerful tool for analyzing protein structure and function. *Fold Des* 1:R121-R126

Hubbard SJ, Thornton JM. NACCESS [Computer Program]. Department of Biochemistry and Molecular Biology, University College London; 1993.

Hwang H, Pierce B, Mintseris J, Janin J, Weng Z (2008) Protein-protein docking benchmark version 3.0. *Proteins* 73:705-709

Innis CA, Anand AP, Sowdhamini R (2004) Prediction of functional sites in proteins using conserved functional group analysis. *J Mol Biol* 337:1053-1068

Ispolatov I, Yuryev A, Mazo I, Maslov S (2005) Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res* 33(11):3629 -3635

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS* 98:4569-4574

Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA (2005) PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res* 33:D183–D187

Izarzugaza JM, Juan D, Pons C, Ranea JA, Valencia A, Pazos F (2006) TSEMA: interactive prediction of protein pairings between interacting families. *Nucleic Acids Res* 34:W315–W319

Jones S, Thornton JM (1996) Principles of Protein-Protein Interactions. *PNAS* 93:13-20

Jones S, Thornton JM (1997) Analysis of Protein-Protein Interaction Sites using Surface Patches. *J Mol Biol* 272:121-132

Juan D, Mellado M, Rodríguez-Frade JM, Hernanz-Falcón P, Serrano A, Del Sol A, Valencia A, Martínez-A C, Rojas AM. (2005) A framework for computational and experimental methods: Identifying dimerization residues in CCR chemokine receptors. *Bioinformatics* 21 Suppl 2:ii13-ii18.

Juan D, Pazos F, Valencia A (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary net- works. *PNAS* 105: 934–939

Kalinina OV, Gelfand MS, Russell RB (2009) Combining specificity determining and conserved residues improves functional site prediction. *BMC Bioinformatics* 10:174-174

Kalinina OV, Novichkov PS, Mironov AA, Gelfand MS, Rakhmaninova AB. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucleic Acids Res* 32:W424-W428.

Kanehisa M, Goto S, Kawashima S, Okuno Y, & Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32:D277-280.

Kass I, Horovitz A (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins* 48:611-617

Keskin O, Ma B, Nussinov R (2005) Hot regions in protein–protein inter- actions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol* 345:1281-1294

Klingström T, Plewczynski D (2010) Protein-protein interaction and pathway databases, a graphical review [Internet]. *Brief. Bioinformatics* Available from: http://www.ncbi.nlm.nih.gov/pubmed/20851835

Koonin EV, Wolf YI, & Karev GP (2002) The structure of the protein universe and genome evolution. Nature 420:218-223.

Kowarsch A, Fuchs A, Frishman D, Pagel P (2010) Correlated Mutations: A Hallmark of Phenotypic Amino Acid Substitutions. *PLoS Comput Biol* 6(9):e1000923

Kramer A, Schneider-Mergener J (1998) Synthesis and screening of peptide libraries on continuous cellulose membrane supports. *Methods Mol Biol* 87:25–39.

Kuriyan J, Eisenberg D (2007) The origin of protein interactions and allostery in colocalization. *Nature* 450(7172):983-990

La D, Sutch B, Livesay DR (2005) Predicting protein functional sites with phylogenetic motifs. *Proteins* 58:309-320

Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R, Cesareni G (2004) Interaction Networks by Proteome Peptide Scanning. *Plos Biol* 2(1):94-103

Landgraf R, Xenarios I, Eisenberg D (2001) Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol* 307:1487-1502.

Larson SM, Di-Nardo AA, Davidson AR (2000) Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J Mol Biol* 303:433-446

Laskowski RA, Chistyakov VV, Thornton JM (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res* 33:D266–268

Lebart L, Morineau A, Warwick KM (1984) Multivariate descriptive statistical analysis. *John Wiley & Sons,* New York p 175

Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* 8(12):995-1005

Lee DA, Rentzsch R, Orengo C (2010) GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. *Nucleic Acids Res* 38(3):720-737

Levy ED (2007) PiQSi: protein quaternary structure investigation. *Structure* 15(11):1364-1367.

Levy ED, Erba EB, Robinson CV, Teichmann SA (2008) Assembly reflects evolution of protein complexes. Nature 453(7199):1262-1265

Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA (2006) 3D complex: a structural classification of protein complexes. *PLoS Comput Biol* 2(11):e155

Li W-H (1980) Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fish. *Genetics* 95:237–258.

Li W, Godzik A (2006) CD-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658-1659

Li X, Keskin O, Ma B, Nussinov R, Liang L (2004) Protein–protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *J Mol Biol* 344:781–795

Lichtarge O, Bourne H, Cohen FE (1996a) Evolutionary conserved Gαβγ binding surfaces support a model of the G protein-receptor complex. *PNAS* 93:7507-7511

Lichtarge O, Bourne HR, & Cohen FE (1996b) An Evolutionary Trace method defines binding surfaces common to protein families. *J Mol Biol* 257:342-358

Lichtarge O, Yao H, Kristensen DM, Madabushi S, Mihalek I (2003) Accurate and scalable identification of functional sites by evolutionary tracing. *J Struct Funct Genomics* 4(2-3):159-166

Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 36:D475-D479

Livesay DR, Jambeck P, Rojnuckarin A, Subramaniam S (2003) Conservation of electrostatic properties within enzyme families and superfamilies. *Biochemistry* 42(12):3464-3473

Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295-299

Lopez D, Pazos F (2009) Gene ontology functional annotations at the structural domain level. *Proteins* 76(3):598-607

Lopez G, Valencia A, & Tress M (2007) FireDB--a database of functionally important residues from proteins of known structure. *Nucleic Acids Res* 35:D219-223

Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302:1401–1404

Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459-473

Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J Mol Biol* 316(1):139-154

Madaoui H, Guerois R (2008) Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *PNAS* 105(22):7708-7713

Manning J, Jefferson E, Barton G (2008) The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. *BMC Bioinformatics* 9(1):51

Marino-Buslje C, Teppa E, Di Doménico T, Delfino JM, Nielsen M (2010) Networks of High Mutual Information Define the Structural Proximity of Catalytic Sites: Implications for Catalytic Residue Identification. *PLoS Comput Biol* 6(11):e1000978

Markova M, Peneff C, Hewlins MJ, Schirmer T, John RA (2005) Determinants of substrate specificity in omega-aminotransferases. *J Biol Chem* 280(43):36409-36416

Marttinen P, Corander J, Törönen P, Holm L (2006) Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics* 22(20):2466 -2474

Maynard Smith J. (1970) Natural Selection and the Concept of a Protein Space. *Nature* 225(5232):563-564.

Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior. *Molecular Biology and Evolution* 21(9):1781 -1791

McCammon JA (1998) Theory of biomolecular recognition. *Curr Opin Struct Biol* 8: 245–249

McCarthy AD, Hardie DG. (1984) Fatty acid synthase: an example of protein evolution by gene fusion. *Trends Biochem Sci* 9:60-63

McDonald AG, Boyce S, Tipton KF (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res* 37:D593-D597

Mewes H, Albermann K, Heumann K, Liebl S, & Pfeiffer F (1997) MIPS: a database for protein sequences, homology data and yeast genome information. Nucl. Acids. Res. 25(1):28-30.

Mihalek I, Res I, Lichtarge O (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J Mol Biol* 336:1265-1282

Mika S, Rost B (2006) Protein–Protein Interactions More Conserved within Species than across Species. *PLoS Comput Biol* 2(7):e79.

Miller I, Miller M (1998) John E. Freund's Mathematical Statistics. *Prentice Hall International.* London

Mirny LA, Gelfand MS (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J Mol Biol* 321:7-20

Mok J, Kim PH, Lam HYK, Piccirillo S, Zhou X, Jeschke GR, Sheridan DL et al. (2010) Deciphering Protein Kinase Specificity Through Large-Scale Analysis of Yeast Phosphorylation Site Motifs. *Sci Signal* 3(109):ra12

Moreira IS, Fernandes PA, Ramos MJ. (2007) Hot spots - A review of the protein-protein interface determinant amino-acid residues. *Proteins* 68(4):803-812

Morillas M, Gómez-Puertas P, Rubí B, Clotet J, Ariño J, Valencia A, Hegardt FG, Serra D, Asins G. (2002) Structural model of a malonyl-CoA-binding site of carnitine octanoyltransferase and carnitine palmitoyltransferase I: mutational analysis of a malonyl-CoA affinity domain. *J. Biol. Chem* 277(13):11473-11480

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, Courcelle E, Das U, Daugherty L, Dibley M, Finn R, Fleischmann W, Gough J,

Haft D, Hulo N, Hunter S, Kahn D, Kanapin A, Kejariwal A, Labarga A, Langendijk-Genevaux PS, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Nikolskaya AN, Orchard S, Orengo C, Petryszak R, Selengut JD, Sigrist CJA, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C. (2007) New developments in the InterPro database. *Nucleic Acids Res* 35:D224-228

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536-540.

Nei M, Roychoudhury AK. (1973) Probability of fixation of nonfunctional genes at duplicate loci. *Am Nat* 107:362–372

Neumann B, Walter T, Hériché J-K, *et al.* Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes (2010). *Nature* 464(7289):721-727

Nimrod G, Glaser F, Steinberg D, Ben-Tal N, Pupko T (2005) In silico identification of functional regions in proteins. *Bioinformatics* 21 Suppl 1:i328-337

Ofran Y, Rost B (2007) Protein–Protein Interaction Hotspots Carved into Sequences. *PLoS Comput Biol* 3(7):e119

Ohno S (1970) Evolution by Gene Duplication. *Springer, New York.*

Ohno S (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Semin Cell Dev Biol* 10(5):517-22

Olmea O, Valencia A (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des* 2: S25-S32.

Onrust R, Herzmark P, Chi P*,* Garcia PD, Lichtarge O, Kingsley C, Bourne HR (1997) Receptor and βγ binding sites in the alpha subunit of the retinal G protein transducin. *Science* 275:381-384

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH--a hierarchic classification of protein domain structures. *Structure* 5(8):1093-1108

Orlowski J, Kaczanowski S, Zielenkiewicz P (2007) Overrepresentation of interactions between homologous proteins in interactomes. *FEBS Letters* 581(1):52-56

Ouzounis C, Perez-Iratxeta C, Sander C, Valencia A. (1998) Are binding residues conserved? *Pacific Symposium on Biocomputing* 3:399-410

Panjkovich A, Aloy P (2010) Predicting protein-protein interaction specificity through the integration of three-dimensional structural information and the evolutionary record of protein domains. *Mol BioSyst* 6(4):741-749

Parthasarathi L, Casey F, Stein A, Aloy P, Shields DC (2008) Approved drug mimics of short peptide ligands from protein interaction motifs. *J Chem Inf Model* 48(10):1943-1948

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. *J Mol Biol* 271(4):511-523.

Pazos F, Ranea JAG, Juan D, Sternberg MJE (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol* 352: 1002–1015

Pazos F, Valencia A (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng* 14: 609–614

Pazos F, Valencia A (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* 47(2):219-227

Pazos F, Valencia A. Protein co-evolution, co-adaptation and interactions. (2008) *EMBO J* 27(20):2648-2655

Pei J, Cai W, Kinch LN, Grishin NV (2006) Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. *Bioinformatics* 22(2):164 -171

Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* 17(8):700-712

Peña D (2002) Análisis de Datos Multivariantes. *McGraw Hill,* Madrid

Pereira-Leal JB, Teichmann SA (2005) Novel specificities emerge by stepwise duplication of functional modules. *Genome Research* 15(4):552 -559

Pierleoni A, Martelli PL, Fariselli P, & Casadio R (2007) eSLDB: eukaryotic subcellular localization database. Nucleic Acids Res 35:D208-212.

Pirovano W, Feenstra KA, Heringa J (2006) Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Res* 34(22):6540-6548

Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res* 32: D129-133.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical Recipes in C: The Art of Scientific Computing. *Cambridge University Press, Cambridge*

Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* 18:S71-S77

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res* 29(1):159-164

Ramani AK, Marcotte EM (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol* 327(1):273-284

Reichmann D, Rahat O, Albeck S, Meged R, Dym O, *et al.* (2005) The modular architecture of protein–protein binding interfaces. *PNAS* 102:57–62

Reineke U, Volkmer-Engert R, Schneider-Mergener J (2001) Applications of peptide arrays prepared by the SPOT-technology. *Curr Opin Biotechnol* 12:59–64

Reva B, Antipin Y, Sander C (2007) Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol* 8(11):R232

Riley M (1993) Functions of the gene products of Escherichia coli. *Microbiol. Rev* 57(4):862-952

Rossmann MG, Argos P (1981) Protein folding. *Annu. Rev. Biochem* 50:497-532

Rost B (2002) Enzyme Function Less Conserved than Anticipated. *J Mol Biol* 318(2):595-608

Sali A (1998) 100,000 protein structures for the biologist. *Nat Struct Biol*; 5: 1029-1032.

Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68

Sangar V, Blankenberg DJ, Altman N, Lesk AM (2007) Quantitative sequence-function relationships in proteins based on gene ontology. *BMC Bioinf* 8:294-294

Sankararaman S, Sjölander K (2008) INTREPID--INformation-theoretic TREe traversal for Protein functional site IDentification. *Bioinformatics* 24(21):2445-52

Sayle R, Milner-White E (1995) RASMOL: biomolecular graphics for all. Trends Biochem Sci 20:374-376

Schueler-Furman O, Baker,D (2003) Conserved residue clustering and protein structure prediction. Proteins 52:225–235

Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. *PNAS* 106(52):22124-22129.

Schultz J, Milpetz F, Bork P, Ponting CP (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *PNAS* 95(11):5857-5864

Schwarz R, Seibel PN, Rahmann S, Schoen C, Huenerberg M, Müller-Reible C, Dandekar T, Karchin R, Schultz J, Müller T. (2009) Detecting species-site dependencies in large multiple sequence alignments. *Nucleic Acids Res* 37(18):5959-5968

Seo J, Lee K-J (2004) Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. *J Biochem Mol Biol* 37(1):35-44

Shi Z, Resing KA, Ahn NG (2006) Networks for the allosteric control of protein kinases. *Curr Opin Struct Biol* 16:686-692

Shoemaker BA, Panchenko AR. (2007) Deciphering Protein–Protein Interactions. Part II. Computational Methods to Predict Protein and Domain Interaction Partners. *PLoS Comput Biol* 3(4):e43

Shou C, Bhardwaj N, Lam HYK, Yan K-K, Kim PM, Snyder M, Gerstein MB (2011) Measuring the Evolutionary Rewiring of Biological Networks. *PLoS Comput Biol* 7(1):e1001050

Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* 3:265-274

Sjölander K (2004) Phylogenomic inferrence of protein molecular function: advances and challenges. Bioinformatics 20(2):170-179.

Snel B, Huynen MA (2004) Quantifying modularity in the evolution of biomolecular systems. *Genome Res* 14(3):391-397

Sonnhammer E, Hollich V. (2005) Scoredist: A simple and robust protein sequence distance estimator. *BMC Bioinformatics* 6:108

Sönnichsen B, Koski LB, Walsh A*, et al.* (2005) Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. *Nature* 434:462-469

Sowa ME, He W, Slep KC, Kercher MA, Lichtarge O, Wensel TG (2001) Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat Struct Mol Biol* 8(3):234-237

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34:D535-539

Stein A, Pache RA, Bernadó P, Pons M, Aloy P (2009) Dynamic interactions of proteins in complex networks: a more structured view. *FEBS Journal* 276(19):5390-5405

Stein A, Russell RB, Aloy P (2005) 3did: interacting protein domains of known three- dimensional structure. *Nucleic Acids Res* 33:D413-D417

Stuart AC, Ilyn VA, Sali A (2002) LigBase: a databaes of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics* 18:200–201

Süel GM, Lockless SW, Wall MA, Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol* 10(1):59-69

Suthram S, Sittler T, Ideker T (2005) The Plasmodium protein network diverges from those of other eukaryotes. *Nature* 438(7064):108-112.

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278(5338):631-637

Thomas PD, Campbell MJ, Kejariwal A, Mi H (2003) Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 13(9):2129-2141

Thorn KS, Bogan AA (2001) ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* 17:284-285

Tian W, Skolnick J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* 333(4):863-882

Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a

structural perspective. *J Mol Biol* 307(4):1113-1143

Tompa P (2003) Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* 25(9):847-855

Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295:321–324

Tonikian R, Xin X, Toret CP, Gfeller D, Landgraf C, Panni S, Paoluzi S, et al. (2009) Bayesian Modeling of the Yeast SH3 Domain Interactome Predicts Spatiotemporal Dynamics of Endocytosis Proteins. *PLoS Biol* 7(10): e1000218. doi:10.1371/journal.pbio.1000218

Tonikian R, Zhang Y, Boone C, Sidhu SS (2007) Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries. *Nat. Protocols* 2(6):1368-1386

Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, Reva B, Held HA, et al. 2008. A Specificity Map for the PDZ Domain Family. *PLoS Biology* 6(9) doi:10.1371/journal.pbio.0060239

Tress M, et al. (2005) Scoring docking models with evolutionary information. Proteins. 60(2):275-280.

Tress ML, Graña O, Valencia A (2004) SQUARE--determining reliable regions in sequence alignments. *Bioinformatics* 20(6):974-995

Triviño JC, Pazos F (2010) Quantitative global studies of reactomes and metabolomes using a vectorial representation of reactions and chemical compounds. *BMC Syst Biol* 4:46

Tsai CJ, Kumar S, Ma B, Nussinov R (1999) Folding funnels, binding funnels, and protein function. *Protein Sci* 8: 1181–1190

Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R. (2009) A survey of available tools and web servers for analysis of protein–protein interactions and interfaces. Briefings in *Bioinformatics* 10(3):217 -232

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. (2000) A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature 403:623-627

Valdar WS (2002) Scoring residue conservation. *Proteins* 48: 227-241

Valdar WS, Thornton JM (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* 42(1):108-124

Valencia A (2005) Automatic annotation of protein function. *Curr Opin Struct Biol* 15(3):267-74

van Dam TJP, Snel B (2008) Protein Complex Evolution Does Not Involve Extensive Network Rewiring. *PLoS Comput Biol* 4(7):e1000132

Vaughan CK, Buckle AM, Fersht AR (1999) Structural response to mutation at a protein–protein interface. *J Mol Biol* 286:1487-1506

Velankar S (2004) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res* 33:D262-D265

Venter JC, Remington K, Heidelberg JF *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66-74.

Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* 18:1283-1292

Wagner A (2003) How the global structure of protein interaction networks evolves. *Proc Biol Sci* 270:457-466

Wagner A (2005) Energy constraints on the evolution of gene expression. *Mol Biol Evol* 22, 1365–1374

Wallace I, Higgins D (2007) Supervised multivariate analysis of sequence groups to identify specificity determining residues *BMC Bioinformatics* 8(1):135

Wass MN, Fuentes G, Pons C, Pazos F, Valencia A (2011) Towards the prediction of protein interaction partners using physical docking. *Mol Syst Biol* doi.org/10.1038/msb.2011.3

Wavreille AS, Pei D (2007) A chemical approach to the identification of tensin-binding proteins. *ACS Chem Biol* 2(2):109–118

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein–protein interaction by message passing. *PNAS* 106(1):67-72

Wennerberg K, Rossman KL, Der CJ (2005) The Ras superfamily at a glance. *J Cell Sci* 118(5):843-846

Wicker N, Perrin GR, Thierry JC, Poch O (2001) Secator: A Program for Inferring Protein Subfamilies from Phylogenetic Trees. *Molecular Biology and Evolution* 18(8):1435-1441

Wilcoxon F(1945) Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1(6):80-83

Wilson D, Madera M, Vogel C, Chothia C, Gough J (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res* 35:D308-313

Winzeler EA, Shoemaker DD, Astromoff A*, et al.* (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* 285:901-906

Wright CF, Teichmann SA, Clarke J, Dobson CM (2005) The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* 438(7069):878-881

Wu CH, Nikolskaya A, Huang H, Yeh L-SL, Natale DA, Vinayaka CR, Hu Z-Z, Mazumder R, Kumar S, Kourtesis P, Ledley RS, Suzek BE, Arminski L, Chen Y, Zhang J, Cardenas JL, Chung S, Castro-Alvear J, Dinkov G, Barker WC (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res* 32:D112-D114

Wu CL, Zukerberg LR, Ngwu C, Harlow E, Lees JA (1995) In vivo association of E2F and DP family proteins. *Mol Cell Biol* 15(5):2536-2546

Wuchty S, Oltvai ZN, Barabási A-L (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet* 35(2):176-179

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D (2000) DIP: the database of interacting proteins. *Nucleic Acids Res* 28(1):289-291

Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* 6(5):1882-1898

Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L, Lichtarge O (2003) An Accurate, Sensitive, and Scalable Method to Identify Functional Sites in Protein Structures. *J Mol Biol* 326(1):255-261

Ye K, Anton Feenstra K, Heringa J, IJzerman AP, Marchiori E (2008) Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. Bioinformatics 24(1):18-25

Yeang CH, Haussler D (2007) Detecting coevolution in and among protein domains. *PLoS Comput Biol* 3(11):e211

Yeats C, Maibaum M, Marsden R, Dibley M, Lee D, Addou S, Orengo CA (2006) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Res* 34:D281-284

Yona G, Linial N, Linial M (1999) ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins.* 37:360-378

Yooseph S, Sutton G, Rusch DB, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biol* 13;5(3):e16.

Zarrinpar A, Bhattacharyya RP, Lim WA (2003) The structure and function of proline recognition domains. Sci STKE 179:RE8.

Zheng N, Fraenkel E, Pabo CO, Pavletich NP (1999). Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP. Genes Dev 13:666-674

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. *J Theor Biol* 8:357-366

Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol* 195:957–961

Anexo I.

Material Suplementario

Anexo I. Material Adicional

A.I. Reproducibilidad de los resultados obtenidos en Resultados Parte II mediante métodos alternativos de detección de subfamilias y SDPs.

En esta sección se pretende comprobar la generalidad/independencia de los resultados obtenidos en el análisis a gran escala presentado en la Parte II de esta tesis (*Estudio a gran escala de la contribución de importantes aspectos de la especificidad funcional a la evolución en secuencia de las familias de proteínas*) respecto del método utilizado para la definición de subfamilias y SDPs. Para ello se reproducen los análisis clave del estudio empleando cuatro métodos representativos del estado actual del campo para los cuales se dispone de software automático: *Evolutionary Trace* (ET; Mihalek *et al.*, 2004), *Combinatorial Entropy Optimization* (CEO; Reva *et al.*, 2007), *Mutational Behavior* (MB; del Sol Mesa *et al.*, 2003) y *Sequence Space* (Casari *et al.*, 1995). En el caso de *Sequence Space*, a la descomposición inicial del MSA mediante PCA descrita por Casari *et al.* se aplica el mismo protocolo que el descrito en la **Sección III.I.1** para S3det, a fin de automatizar la detección de subfamilias y SDPs, las cuales -en el artículo original- se definen manualmente. Por esta razón se referirá a esta implementación como *SequenceSpace*^{S3det} para distinguirla del original. En esencia, la diferencia aquí entre S3det y *SequenceSpace*^{S3det} se limita a que el primero parte de un tratamiento con MCA y el segundo con PCA.

Método	Categoría Metodológica	Detección de Subfamilias	Umbrales automáticos para la definición de SDPs	Referencia
S3det	Análisis Multivariante (MCA)	**Sí**	**Sí**	Rausell *et al.* PNAS (2010) y este trabajo
SequenceSpace^{S3det}	Análisis Multivariante (PCA)	**Sí**	**Sí**	Casari *et al. Nat Struct Biol* (1995) y este trabajo
MB	Correlación entre la variabilidad dentro de posición y entre las secuencias	**No**	**Sí**	del Sol *et al. J Mol Biol* (2003)
ET	Asistido por árboles filogenéticos y basado en entropía	**No**	**No**	Mihalek *et al. J Mol Biol* (2004)
CEO	Basado en entropía	**Sí**	**No**	Reva *et al. Genom Biol* (2007)

Tabla S1. Características principales de los métodos ensayados para la definición de subfamilias y SDPs en familias de proteínas (ver texto)**.**

Todos estos métodos se basan en información de secuencia (esto es, parten de un MSA y no requieren información estructural) y son no supervisados (es decir, no necesitan una clasificación funcional establecida *a priori*). En la

tabla S1 se resumen sus principales características. Estos métodos cubren los principales abordajes desarrollados para la detección de SDPs (**Tabla 1**).

Los métodos anteriores se aplicaron a la misma colección de familias Pfam descrita en **Métodos Sección VI.I**, obteniendo los correspondientes conjuntos de subfamilias y SDPs según el caso.

Se estudió en primer lugar la correspondencia entre las subfamilias arrojadas por los métodos *SequenceSpace*^{S3det} y CEO con grupos diferenciales de EC e interactores específicos, de la misma forma que se hizo para el análisis mostrado en **Resultados Sección III.II.1**. En la **Fig. S1** se muestran los resultados obtenidos en forma de histogramas bidimensionales sobre el espacio ROC de las familias Pfam según sus valores de sensibilidad y especificidad respecto a las etiquetas funcionales correspondientes.

La comparación de estos espacios con los obtenidos en el caso de S3det (**Fig. 12**) soporta el acuerdo general de las subfamilias tanto con grupos diferenciales de EC como con grupos de interactores específicos. No obstante esta tendencia global, se pueden observar diferencias menores entre los métodos. CEO tiende a ser más específico que S3det y *SequenceSpace*^{S3det}, si bien a costa de ser menos sensible. Resulta interesante señalar que, en conjunto, las subfamilias provistas por S3det se comportan mejor de forma cualitativa de acuerdo a ambas clasificaciones funcionales estudiadas.

En segundo lugar, se realizó el análisis del enriquecimiento en sitios funcionales de las SDPs detectadas por los cuatro métodos mencionados de forma análoga al presentado en las **Secciones III.III.2** y **III.III.4**. Los softwares de ET y CEO no proveen un umbral (*cutoff*) para sus puntuaciones (*scores*) a partir del cual definir el conjunto de SDPs predicho. En su lugar, la salida de los programas ofrece un ranking de todas las posiciones del MSA. Para hacer comparables los métodos entre sí, en estos dos métodos se seleccionaron las primeras “n” posiciones para cada familia, donde “n” es el número medio de SDPs predicho por S3det, *SequenceSpace*^{S3det} y MB. A falta de un mejor criterio, se pretende así que el número de SDPs entre métodos sea de orden similar, si bien esta forma de proceder podría no ser óptima para ET y/o CEO.

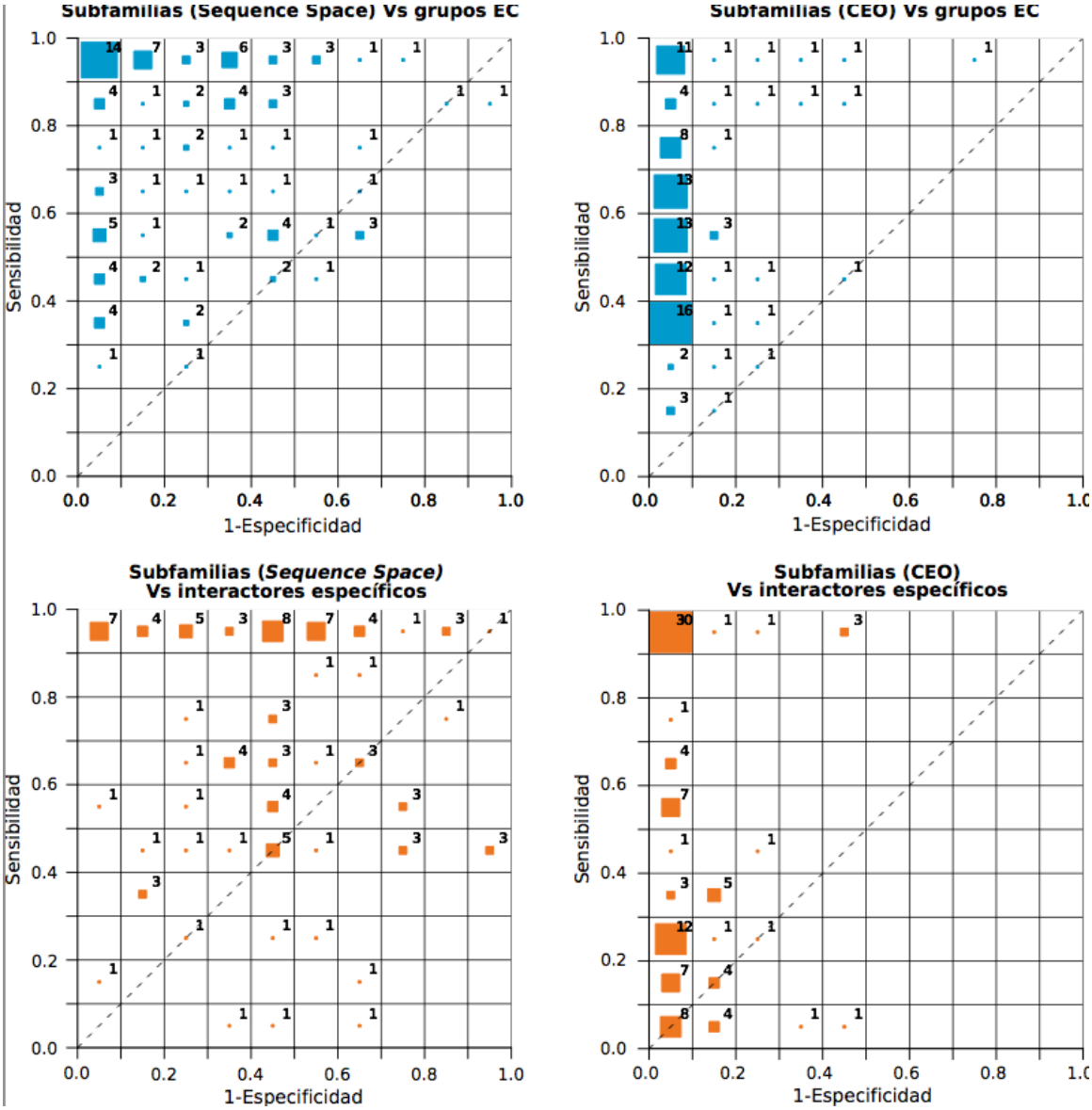


Fig. S1. Correspondencia entre las diferentes subfamilias obtenidas mediante *Sequence Space* (izquierda) y CEO (derecha), y los grupos de ECs (arriba) e interactores específicos (abajo). Los valores de sensibilidad y especificidad obtenidos para cada una de las familias Pfam estudiadas se representan en sendos espacios ROC en los que la distribución de familias se muestra en forma de histograma bidimensional. En cada región discreta se representa el porcentaje de familias con valores de sensibilidad y sensibilidad dentro del intervalo correspondiente. Los valores porcentuales se han redondeado al entero más cercano en aras de la simplicidad (por lo que su suma puede ser diferente de cien) y su representación gráfica se acompaña mediante cuadrados coloreados cuyo lado es proporcional al valor correspondiente.

Los p-valores de los test de enriquecimiento obtenidos para cada uno de los métodos ensayados se muestran en la **Tabla S2**.

p-valores *(sobre el total)*	Total residuos funcionales	Sitio de unión a ligando	Interfaz total	HETERO	HOMO	INTRA
Conservación	**1.92E-27**	**4.73E-20**	**1.92E-09**	**5.32E-02**	**9.09E-03**	**4.29E-07**
ET	**5.55E-26**	**2.85E-22**	**4.87E-11**	**2.86E-04**	**1.79E-02**	**3.59E-07**
ET-*Consfree*	**7.17E-15**	**7.04E-13**	**4.85E-06**	**8.24E-04**	**7.09E-02**	**3.23E-03**
S3det	**1.67E-05**	**4.25E-04**	**1.89E-02**	**1.75E-02**	5.13E-01	4.99E-01
***SeqSpace*^{S3de}**	**3.91E-05**	**3.82E-03**	**2.59E-02**	**7.50E-02**	5.66E-01	4.51E-01
MB	**5.59E-06**	**1.63E-02**	**3.03E-04**	1.79E-01	1.40E-01	**4.32E-02**
CEO	8.27E-02	7.64E-01	1.64E-01	3.05E-01	5.02E-01	4.32E-01

Diferencia de medianas *(sobre el total)*	Total residuos funcionales	Sitio de unión a ligando	Interfaz total	HETERO	HOMO	INTRA
Conservación	**15.25%**	**14.60%**	**6.34%**	**1.84%**	**1.99%**	**6.76%**
ET	**13.49%**	**13.72%**	**6.12%**	**4.82%**	**1.69%**	**5.56%**
ET-*Consfree*	**8.54%**	**7.94%**	**3.74%**	**4.08%**	**1.21%**	**2.30%**
S3det	**5.32%**	**3.97%**	**2.00%**	**4.77%**	-0.01%	0.00%
***SeqSpace*^{S3det}**	**4.48%**	**2.80%**	**1.65%**	**1.94%**	-0.09%	0.17%
MB	**3.62%**	**1.73%**	**2.54%**	1.33%	0.81%	**1.37%**
CEO	1.29%	-0.56%	0.85%	0.66%	0.00%	-0.71%

Tabla S2. Resultados de los test de Wilcoxon de suma de rangos evaluando el enriquecimiento en diferentes tipos de regiones funcionales de SDPs definidos según diferentes metodologías. A fin de favorecer la comparación de sus valores, se incluyen los valores oportunos de la **Tabla 5** correspondientes a las posiciones conservadas y a las SDPs obtenidas por S3det.

La asociación estadística general entre SDPs tanto con sitios de unión a ligando como con interfaces aparece mediante todos los métodos ensayados, a excepción del método CEO. El criterio utilizado para seleccionar un número equivalente de SDPs en este método (ver arriba) podría estar perturbando la bondad de sus resultados. Sin embargo, un análisis en mayor detalle del funcionamiento de CEO está fuera del ámbito de esta tesis.

Cabe destacar el elevado enriquecimiento que muestran los resultados de ET en todos los tipos de residuos funcionales, a unos niveles similares a los de las posiciones conservadas. De hecho, el 76% de las predicciones arrojadas por ET presentan un nivel de conservación igual o superior al 90% en identidad de secuencia, criterio utilizado en este trabajo para definir el conjunto de posiciones conservadas. En contraste, las todos los otros métodos presentan menos del 3% de posiciones conservadas en sus predicciones. Estos resultados son coherentes con la definición más amplia -respecto del resto de métodos- que en ET se hace de las “posiciones relevantes evolutivamente” y que incluye a las posiciones totalmente conservadas.

Para soslayar el elevado solape entre las predicciones de ET y la conservación, se filtraron las posiciones conservadas de los rankings provistos

por este método, generando así un nuevo conjunto de predicciones para cada familia. A los resultados así obtenidos se les identifica aquí como ET-*ConsFree* (resultados de *Evolutionary Trace* sin posiciones conservadas al 90 % de identidad de secuencia). A medida que se filtra la conservación, los resultados de ET devienen gradualmente más parecidos a los del resto de métodos.

Tomados en conjunto, los resultados obtenidos para S3det, *SequenceSpace*^{S3det}, MB y ET-*ConsFree* soportan la asociación general de SDPs tanto con sitios de unión a ligando como con interfaces. Además, las tendencias de enriquecimiento observadas en la **Tabla S2** apuntan a una asociación consistente entre SDPs e interfaces heterodiméricas. Sin embargo, en el caso de interfaces homoméricas e intracadena, los resultados continúan sin ser concluyentes.

S3det produce unos resultados cualitativamente similares a los producidos por los otros métodos. Sin embargo, la capacidad de S3det de predecir subfamilias y SDPs de forma simultánea bajo un mismo marco metodológico es especialmente adecuada para el estudio a gran escala propuesto en la parte II de esta tesis.

Anexo II.

Copia de los artículos publicados por el doctorando relacionados con la tesis

Protein interactions and ligand binding: From protein subfamilies to functional specificity

Antonio Rausell[a], David Juan[a], Florencio Pazos[b], and Alfonso Valencia[a,1]

[a]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/ Melchor Fernández Almagro 3, 28029 Madrid, Spain; [b]Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), C/ Darwin 3, Cantoblanco, 28049 Madrid, Spain

Edited by Barry H. Honig, Columbia University / HHMI, New York, NY, and approved November 27, 2009 (received for review July 20, 2009)

The divergence accumulated during the evolution of protein families translates into their internal organization as subfamilies, and it is directly reflected in the characteristic patterns of differentially conserved residues. These specifically conserved positions in protein subfamilies are known as "specificity determining positions" (SDPs). Previous studies have limited their analysis to the study of the relationship between these positions and ligand-binding specificity, demonstrating significant yet limited predictive capacity. We have systematically extended this observation to include the role of differential protein interactions in the segregation of protein subfamilies and explored in detail the structural distribution of SDPs at protein interfaces. Our results show the extensive influence of protein interactions in the evolution of protein families and the widespread association of SDPs with protein interfaces. The combined analysis of SDPs in interfaces and ligand-binding sites provides a more complete picture of the organization of protein families, constituting the necessary framework for a large scale analysis of the evolution of protein function.

functional residues | protein family evolution | protein function | protein–protein interfaces | specificity determining positions

The structure of protein families is shaped by the sequence divergence accumulated as a consequence of speciation, gene duplication, and deletion events, as well as by the evolutionary selective pressure exerted on each protein in accordance with the corresponding 3D structure and the specific function performed (1, 2). The balance between genomic rearrangements and selective pressure to increase the functional repertoire available to organisms leads to the appearance of new subfamilies in evolutionary time (3).

There are many aspects of protein function that contribute to the evolution of the family organization. These may include the global conservation of catalytic mechanisms (in the case of enzymes), specific binding to substrates and cofactors, as well as the interaction with other proteins in processes such as cell signaling, the regulation of reactions and the formation of macromolecular complexes. Interestingly, even though specific protein interactions certainly are an important part of protein function, the organization of protein families in relation to the specific interactions of different subfamilies remains a poorly studied aspect of functional specificity.

Multiple sequences alignments (MSAs) provide essential information on the evolution of protein families. The positions in MSAs can be interpreted in terms of the amino acid changes allowed or disallowed during evolution, and therefore useful information at the residue level can be inferred from them (4). The most obvious example is the study of fully conserved positions that pinpoint important residues for the structure and function of the family members (5).

A subtler pattern of conservation is represented by the positions differentially conserved within subfamilies. A commonly accepted working hypothesis is that whereas fully conserved positions are related to functional features common to all the members of the family, these other residues are related to functional specificity (e.g., binding of different cofactors). For this reason, they have been termed "specificity determining positions" (SDPs). A variety of computational methods have been used to detect conserved positions and SDPs in MSAs (6–12); for a review see ref. 13. Moreover, the implication of SDPs in determining the differential binding to substrates and interaction partners has been experimentally followed up in a number of cases (14–16).

Despite these efforts, fundamental questions regarding the association between subfamilies, SDPs, and function remain largely unexplored at the systematic level. Notwithstanding, the information currently available on protein sequences, structures, functions, and interactions opens the door to performing more comprehensive studies of the relationships between family organization and functional divergence (17). Indeed, such studies can involve biochemical function and protein interaction specificity. Similarly, they can take into account the associated conservation at the molecular signatures level (SDPs) in fundamental regions corresponding to ligand-binding sites and protein interaction sites.

To carry out a unified analysis of subfamilies and associated SDPs, we have developed a protocol based on multiple correspondence analysis (MCA) (18) that can detect both entities simultaneously. Here we apply this methodology to the largest possible dataset of eukaryotic protein families for which it was possible to compile reliable information on catalytic activity, ligand binding, and protein interactions. The results are interpreted in terms of the relationship between the internal structure of protein families, their functional properties, and specific molecular signatures, with particular attention to the analysis of protein interaction sites.

Results

Functions in Protein Families: Biochemical and Protein Interaction Specificity. This work evaluates the influence of functional constraints on protein family evolution by studying the functional features associated with their subfamily organization and their corresponding SDPs. For this purpose, we developed a multivariate-based protocol capable of detecting protein subfamilies and SDPs in a concomitant way and applied it to a collection of eukaryotic Pfam families (see *Methods*). In this section, the internal organization of protein families in subfamilies was analyzed on a collection of cases for which functional information is available regarding (*i*) their catalytic mechanism as defined in the Enzyme Commission (EC) classes, and (*ii*) the specificity of protein interactors for *Saccharomyces cerevisiae* and *Homo sapiens*, as inferred from "small scale" experiments, which provides a sound basis for the definition of interaction specificities within protein families (see *Methods*).

When 149 families with a representative number of EC labels and 72 families with a representative number of identified

Author contributions: A.R., D.J., and A.V. designed research; A.R. performed research; A.R. and D.J. contributed new reagents/analytic tools; and A.R., D.J., F.P., and A.V. analyzed data and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: valencia@cnio.es.

This article contains supporting information online at www.pnas.org/cgi/content/full/0908044107/DCSupplemental.

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 12 2006, pages 1440–1448
doi:10.1093/bioinformatics/btl104

Sequence analysis

Phylogeny-independent detection of functional residues

Florencio Pazos*, Antonio Rausell and Alfonso Valencia

Protein Design Group, National Centre for Biotechnology (CNB-CSIC), C/Darwin, 3. Campus U. Autónoma, 28049 Cantoblanco, Madrid, Spain

Received on January 19, 2006; revised and accepted on March 16, 2006
Advance Access publication March 21, 2006
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Current projects for the massive characterization of proteomes are generating protein sequences and structures with unknown function. The difficulty of experimentally determining functionally important sites calls for the development of computational methods. The first techniques, based on the search for fully conserved positions in multiple sequence alignments (MSAs), were followed by methods for locating family-dependent conserved positions. These rely on the functional classification implicit in the alignment for locating these positions related with functional specificity. The next obvious step, still scarcely explored, is to detect these positions using a functional classification different from the one implicit in the sequence relationships between the proteins. Here, we present two new methods for locating functional positions which can incorporate an arbitrary external functional classification which may or may not coincide with the one implicit in the MSA. The *Xdet* method is able to use a functional classification with an associated hierarchy or similarity between functions to locate positions related to that classification. The *MCdet* method uses multivariate statistical analysis to locate positions responsible for each one of the functions within a multifunctional family.

Results: We applied the methods to different cases, illustrating scenarios where there is a disagreement between the functional and the phylogenetic relationships, and demonstrated their usefulness for the phylogeny-independent prediction of functional positions.

Availability: All computer programs and datasets used in this work are available from the authors for academic use.

Contact: pazos@cnb.uam.es

Supplementary information: Supplementary data are available at http://pdg.cnb.uam.es/pazos/Xdet_MCdet_Add/

INTRODUCTION

If the genomic era was characterized by the massive sequencing of complete genomes, the so-called 'post-genomic' era is being, may be, characterized by an unexpected lack of tools for obtaining relevant information from these raw sequences. Today, we know the complete sequences of hundreds of genomes from the three kingdoms, and 'environmental sequencing' (Venter *et al*., 2004) (the organism-independent sequencing of DNA repertories directly extracted from environmental samples) is boosting the number of available sequences. There is also an increasing number of proteins of known three-dimensional (3D) structures without associated functional information, in part owing to the Structural Genomics projects (Brenner, 2001).

Determining which residues in a protein are responsible for its function is very important in order to understand its molecular mechanism, to modify this function in our benefit (biotechnology) or to correct problems related with this function (e.g. pathologies). The experimental characterization of function and functional features (functional sites, etc.) is very expensive, time consuming and difficult to automate. This justifies the development of computational methods for predicting functional sites and other functional features for these uncharacterized sequences.

Some methods use previously known functional sites to derive sequence profiles (Mulder *et al*., 2003) or structural templates (Di Gennaro *et al*., 2001; Porter *et al*., 2004) in order to match new sequences/structures against them. Other techniques are able to detect functional sites without previous knowledge of them. Some of these methods are able to predict functional sites based on single sequences, like the method developed by Ofran and Rost for predicting protein interaction sites (Ofran and Rost, 2003). Others are based on single 3D structures. They look for structural features frequently associated with active sites and binding sites, like low-stability regions (Elcock, 2001) or special connectivity patterns extracted from residue–residue contact networks (Del Sol and O'Meara, 2004). Nevertheless, most of the methods predict functional positions based on multiple sequence or structure alignments of related proteins, and they work under the common assumption of conservation of functional residues during evolution.

The advantage of structural alignments is that they can relate remote homologs (Pazos and Sternberg, 2004), and their drawback is that they need 3D structures to work, which are more scarce than sequences.

Since sequences are still more abundant than structures, there is a plethora of methods for predicting functional sites from sequence alignments. The first information extracted from sequence alignments was related with fully conserved positions (Zuckerkandl and Pauling, 1965). Fully conserved positions are related with sites important for the function or the structure of the protein. Later, the concept of conservation was extended to family-dependent conservation: positions that are conserved within subfamilies being the aminoacid type different between different subfamilies. These

*To whom correspondence should be addressed.